Perceptual techniques in audio quality assessment

Antony W. Rix

A thesis submitted for the degree of Doctor of Philosophy

University of Edinburgh

2003

Abstract

This thesis discusses quality assessment of audio communications systems, in particular telephone networks. The goal is a model to estimate perceived quality, measured by subjective tests. Several authors have described perceptual models for this purpose, based on comparison of auditory transforms, which have been found to generalise better than simpler objective measures such as signal-to-noise ratio. However, it is shown that these early perceptual models give inaccurate scores for end-to-end measurements of networks that may include filtering or variable delay. These problems are investigated and addressed in this thesis.

Perceptual models compare transforms of the input and output signals, requiring the time-delay to be estimated. Previous methods are found to fail with some types of linear and non-linear distortion, or are prohibitively slow. A new technique for time-delay estimation based on a smoothed weighted histogram of frame-by-frame delays is presented. This has low complexity and is found to be more robust to non-linear distortions typical of telephone networks. This technique is further extended to identify piecewise constant delay, enabling models to be used for assessing packet-based transmission such as voice over IP, where delay may change several times during a measurement.

Prior speech quality measurement models did not distinguish non-linear processing and linear filtering. The latter is normally less subjectively disturbing, and it is shown that equalisation improves the accuracy of perceptual models for measurements that may include analogue or acoustic components. Linear transfer function estimation is found to be unreliable due to non-linear distortions. Spectral difference and phaseless cross-spectrum estimation methods for identifying and equalising the linear transfer function are implemented for this application, operating in the filterbank and short-term Fourier spectrum domains.

This thesis provides the first detailed examination of the process of selecting and mapping multiple objective perceptual distortion parameters to estimated subjective quality. The systematic variation of subjective opinion between tests is examined and addressed using a new method of monotonic polynomial regression. The effect on conventional regression techniques, and a new joint optimisation process, are considered. Novel techniques for monotonic regression, preserving knowledge of the sign of the underlying relationship, are introduced. These are applied to a new perceptual model and performance is analysed over a large database of subjective tests.

The results show that perceptual models that use the techniques introduced in this thesis can give accurate results, both for end-to-end measurement of telephone networks and for their components such as speech coders, with most currently available technologies.

Declaration of originality

I hereby declare that

(i) except where indicated in the text, the research recorded in this thesis was conducted entirely by myself; and

(ii) this thesis was composed and originated entirely by myself; and

(iii) this work has not been submitted for any other degree or professional qualification.

Antony Rix

Acknowledgements

The bulk of this study was funded by an Industrial Fellowship from the Royal Commission for the Exhibition of 1851. BT Laboratories sponsored the first year, and Psytechnics has supported (and distracted) me during writing up.

I must acknowledge the technical contribution of Mike Hollier, Richard Reynolds and Phil Gray to this work. Mike has been a source of challenge, encouragement and guidance as industrial supervisor and manager throughout this study. Discussions with him and with Phil and Richard fostered the early development of PAMS, in particular the histogram-based delay estimation and segmental utterance processing.

Eleanor Beamond worked under my direction to explore time alignment and gradient descent methods, and helped to formulate an early version of the utterance splitting algorithm.

I have also enjoyed collaborating with John Beerends and Andries Hekstra to implement the ideas described in this thesis in PESQ, and with Tom Goldstein, John Beerends and Jens Berger on P.AAM.

Many other people have provided subjective test data, tools or other assistance, including Paul Barrett, Ludovic Malfait and Andrew Whitefield at Psytechnics, and Philip Arden, David Hands and Rupert Voelcker at BT. I have also received data from the other proponents in the ITU-T study group 12 competitions for P.862, P.SEAM and P.AAM.

At the University of Edinburgh, Mervyn Jack made sure that I was aware of the submission deadline, and Steve McLaughlin gave valuable help with the planning of this thesis and with comments on earlier drafts.

I would like to thank my parents and brother, who lent support and advice. And most of all, I want to thank my wife, Kathryn, for her love, motivation and proofreading, and for being there.

Contents

List of figures

List of tables

Acronyms

2G	Second-generation mobile network
3G	Third-generation mobile network
ACR	Absolute category rating [ITU-T P.800]
ADPCM	Adaptive delta pulse code modulation [ITU-T G.726]
AES	Audio Engineering Society
AMR	Adaptive multi-rate [GSM 06.90]
ASD	Auditory spectrum difference [Karjalainen 1985]
ATM	Asynchronous transfer mode
BSD	Bark spectral distortion [Wang 1992]
CCR	Comparison category rating
CELP	Codebook excited linear predictor
DAM	Diagnostic acceptability measure [Voiers 1977]
dBov	Decibels with respect to peak overload point
DCME	Digital circuit multiplication equipment
DCR	Degradation category rating
DCT	Discrete cosine transform
DFT	Discrete Fourier transform
DIX	Distortion index [Thiede 1996]
DTMF	Discrete tone multi-frequency
DTW	Dynamic time-warping
DTX	Discontinuous transmission
EFR	Enhanced full-rate [GSM 06.60]
ERB	Equivalent rectangular bandwidth
ETSI	European Telecommunications Standards Institute
EVRC	Enhanced variable-rate coder [TIA/EIA IS-127]
FDM	Frequency division multiplexing
FER	Frame erasure
FFT	Fast Fourier transform
FIR	Finite impulse response
FR	Full-rate [GSM 06.10]
GSM	Global system for mobile
HATS	Head and torso simulator [ITU-T P.58]
HR	Half-rate

ICASSP	International conference on acoustics, speech and signal processing
IEC	International Electrotechnical Commission
IEE	Institution of Electrical Engineers
IEEE	Institute of Electrical and Electronics Engineers
IID	Independent, identically distributed
IIR	Infinite impulse response
IP	Internet protocol
IRS	Intermediate reference system [ITU-T P.48]
ISO	International Standards Organization
ITU-R	International Telecommunication Union - Radiocommunication bureau
ITU-T	International Telecommunication Union - Telecommunications standardisation bureau
LE	Listening effort
LPC	Linear predictive coding
LQ	Listening quality [ITU-T P.800]
LTI	Linear, time invariant
MAF	Minimum audible field
MAP	Maximum a posteriori
MARS	Multivariate adaptive regression splines [Friedman 1991]
MDCT	Modified discrete cosine transform
MELP	Mixed-excitation linear predictor
MIRS	Modified intermediate reference system [ITU-T P.830]
ML	Maximum likelihood
MLP	Multi-layer perceptron
MNB	Measuring normalising blocks [Voran 1999a, ITU-T P.861].
MNRU	Modulated-noise reference unit [ITU-T P.810]
MOS	Mean opinion score [ITU-T P.800]
MOV	Model output variable [ITU-R BS.1387]
MPEG	Motion picture experts group [Brandenburg 1996]
MSE	Mean squared error
MSIN	Mobile station input filter
MUSIC	Multiple signal classification
NLMS	Normalised least mean squares
ODG	Objective difference grade [ITU-R BS.1387]
OSQ	Objective speech quality
P.AAM	Perceptual acoustic assessment measure
PAMS	Perceptual analysis measurement system [Rix 2000b]
PAQM	Perceptual audio quality measure [Beerends 1992]

PC	Personal computer
PCA	Principal component analysis
PCM	Pulse code modulation [ITU-T G.711]
PDF	Probability density function
PEAQ	Perceptual evaluation of audio quality [ITU-R BS.1387]
PESQ	Perceptual evaluation of speech quality [ITU-T P.862, Rix 2002b, Beerends 2002].
PSQM	Perceptual speech quality measure [Beerends 1994, ITU-T P.861]
PSTN	Public switched telephone network
RMS	Root mean squared
RMSE	Root mean squared error
RPE-LTP	Residual pulse excitation-linear transform predictor
SDG	Subjective difference grade [ITU-R BS.1116]
SDH	Synchronous digital hierarchy
SIAM	Society for Industrial and Applied Mathematics
SNR	Signal-to-noise ratio
SPL	Sound pressure level
SSB	Single side-band
STFT	Short-term Fourier transform
TFE	Transfer function estimate
TIA	Telecommunications Industry Association
TOSQA	Telecommunication objective speech quality assessment [Berger 1997]
VAD	Voice activity detection
VoIP	Voice over Internet protocol
VXC	Vector excitation coding

Symbols

Symbol	Meaning
α	Histogram weight power
α_m	Lower bound of parameter x_m
β	Histogram smoothing kernel width
β_m	Upper bound of parameter x_m
γ	Zwicker sone exponent [Zwicker 1990]
γ_1, γ_2	Coefficients of linear mapping on MLP output
$\delta(t)$	Discrete-time Dirac delta function (unit impulse)
ε	Prediction error $\hat{y} - y$
$\Theta(\,)$	Order of computational complexity within a scale and constant offset
$\varkappa(\tau)$	Smoothing kernel at lag τ
λ	Exponent in computation of modified local frame coherence
ν_1, ν_2	Bounds on a parameter
ϱ	Correlation coefficient
$\tau(t)$	Time-delay between $r(t)$ and $d(t)$ at time t ($\tilde{\tau}$ represents estimated delay)
$\Phi_{rd}(\Omega)$	Phaseless cross-spectrum estimate
$\chi(\tau,k)$	Cross-correlation function between $r(t)$ and $d(t)$ in frame k with lag τ
$\psi(\Omega)$	Generalised cross-correlator weight function [Knapp 1976]
Ω	Radian frequency
a, a_1, a_2, a_3, b, c, d	General unknown coefficients; coefficients of probability density functions
a, **b**	Coefficient vectors
C	Regression weighted cost
$c_l(k)$	Modified local frame coherence
$C_{rd}(\Omega)$	Magnitude squared coherence
$d(t)$	Degraded signal
$d_s(k,f)$	Degraded sensation (excitation) surface
D	Severe distortion event ($\tilde{D}$ represents the absence of distortion)
e	exp(1)
$e_s(k,f)$	Error surface $d_s(k,f)-r_s(k,f)$
E[ ]	Expectation operator
F[ ]	Fourier transform operator ($F^{-1}[\,]$ is the inverse Fourier transform)
f	Frequency index
$f(\,)$	Mapping function from distortion parameters to OSQ
$g(\,)$	Mapping function from OSQ to MOS

Symbol	Meaning
$h(t)$	Linear system impulse response ($\hat{h}(t)$ estimated impulse response, etc.)
$H(z)$	Two-sized z-transform of $h(t)$: $H(z) = \sum_{t=-\infty}^{\infty} h(t)z^{-t}$
$H(e^{j\Omega})$	Discrete Fourier transform of $h(t)$, etc.
H	Number of nodes in MLP hidden layer (Chapter 5)
i	General index variable
j	$\sqrt{-1}$
k	Frame index (Chapters 2–4); condition number (Chapter 5)
K	Total number of conditions (Chapter 5)
K_{eff}	Estimated effective total number of conditions (Chapter 5)
log( )	Logarithm to the base e
$L(\ )$	Log-probability or log-likelihood function
m	Distortion parameter index; mapping function or polynomial order
M	Parameter subset size
$n(t)$	Noise signal
N_f, N_k, N_r	Number of points on FFT, number of conditions k, number of samples in $r(t)$, etc.
p	Lebesgue L_p-norm power [Quackenbush 1988]
$\mathbf{p}$	Bezier spline coefficient vector
$p(\ )$	Probability density function
$p_h(\tau)$	Weighted delay histogram
$p_s(\tau)$	Smoothed weighted delay histogram
$P_{rd}(\Omega)$	Cross-spectrum between r and d at radian frequency Ω. Similarly $P_{rn}(\Omega)$, etc.
$P_{rr}(\Omega)$	Auto-spectrum of signal r at radian frequency Ω. Similarly $P_{dd}(\Omega)$, $P_{nn}(\Omega)$, etc.
Q	MNRU SNR, dB
$r(t)$	Reference signal
$r_s(k,f)$	Reference sensation (excitation) surface
s	Subjective test index
t	Discrete-time index
$w(s)$	Subjective test weight (Chapter 5)
$w(t)$	Window function (Chapters 3, 4)
W_{sil}	PSQM silent interval weighting
$x(i)$	Distortion parameter
y	Subjective MOS
$\hat{y} = g(y_0, \mathbf{b})$	y_0 mapped with minimum MSE to y for a given subjective test
y_0	Objective speech quality score
y_{LE}, y_{LQ}	PAMS listening effort, listening quality
z^{-1}	Unit delay operator

CHAPTER 1

Introduction

1.1 Overview

This thesis addresses methods for automatic estimation of the perceived quality of audio transmission systems based on intrusive measurements. The main focus is the speech quality of telecommunications networks. In other words, after making a test call or calls, a quality score is computed which maps directly to an easily-understood subjective opinion scale such as *excellent, good, fair, poor, bad.*

The development of these measurement methods has been driven by the growing complexity of the systems that they are required to test – for telecommunications, these may include low bit-rate coders, error-prone channels, linear filtering, noise, and other forms of signal processing such as noise reduction. The potential non-linearity of these systems means that conventional signal processing measures have very limited applicability. This has motivated the development of perceptual models, which process signals using transforms that are based on psychoacoustic principles. To calibrate and evaluate these models requires subjective tests, as the objective is to predict what an end-user would think of the quality. Thus this study draws on the fields of signal processing, communications, psychoacoustics and subjective testing.

Before this study, a number of perceptual models had been developed for testing relatively simple systems such as low bit-rate coders. These were generally single-parameter models, which essentially computed the average audible difference between the input and output of the system. It was found that these simple models had limitations which made them inaccurate for testing real-world systems such as telephone networks.

These problems were due to three main issues that had been largely neglected in the development of earlier perceptual models: identification of time-delay, linear filtering in the system under test, and the multi-dimensional nature of subjective quality perception.

Time-delay. The basic cross-correlation method of delay identification, used by most other authors, can lead to inaccurate results with highly non-linear processing or delay variations.

Linear filtering is common in analogue or acoustic interfaces and is perceived very differently from non-linear coding distortions – but this was not accounted for by previous models.

Subjective perception. The complex nature of the subjective quality judgement requires multi-parameter regression, incorporating variations between subjective tests, to develop more accurate and robust models.

The solution of these three issues forms the core of this thesis.

By characterising these problems and identifying solutions to them, this work has made a key contribution to the field of perceptual quality measurement. Two models that apply the new techniques presented in this thesis, PAMS and PESQ, are now in use by thousands of telecommunications engineers world-wide, and PESQ has become the main international standard in this area.

1.2 Previous work

The main background to this work centres on the perceptual models of Hollier [Hollier 1995] and Beerends [Beerends 1994, ITU-T P.861]. This section considers how and why these models were developed and provides a brief introduction to this field. Further details are set out in chapter 2.

During the 1970s it became apparent that it was advantageous to use properties of the human auditory system to improve audio coders, in particular speech coders. By then, results were available on how people perceive properties such as loudness, pitch and timbre, and how differences between signals may be masked in time and/or frequency. In addition, computational power had developed to the point where these properties could be practically modelled. In speech and audio coders, greater quality can be delivered by choosing the encoded coefficients so as to minimise the audibility of the resultant distortions, rather than a simpler metric such as mean squared error. This improvement was first introduced for speech coding by Schroeder et al [Schroeder 1979], and is now in widespread use in many speech and audio coders.

Perceptual coding is one example of non-linear processing that may be encountered in audio communications systems. Other examples include error processes, discontinuous transmission with comfort noise insertion during silent periods, and noise reduction. In all of these cases the overall perceived quality is affected in a complex way, and is not accurately modelled by simple objective measures such as root mean squared error (RMSE), noise level, or frequency response. For example, if white noise is added to speech or music at a signal-to-noise ratio (SNR) of 13dB, it is clearly audible and very disturbing. However, if the noise is shaped in time and frequency so as to be below the masked threshold, it is almost impossible for a human to

detect, even in the most stringent listening conditions. Yet in these two cases, the noise level, SNR, RMSE and linear frequency response of the system under test are identical.

Because of this, subjective testing became the key measure of quality of these systems. In a typical listening test, subjects listen to recordings that have been processed through a range of conditions, and vote on a simple opinion scale such as that shown in Table 1.1 [ITU-T P.800, ITU-T P.830, ITU-R BS.1116]. The average score for a condition, across all subjects, is known as the mean opinion score (MOS).

Table 1.1: Listening quality opinion scale

Quality of the speech	
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Subjective testing does, however, have a number of limitations. Tests must be run in special noise-controlled listening rooms, using professional equipment, and are time-consuming. This makes them expensive and slow to conduct, and limits the number of conditions that can be tested. It is also difficult to obtain fully repeatable results, as different subjects have different assumptions and opinions. A computational model that gives an accurate prediction of subjective quality would be much more suitable for field applications.

To achieve this, it was found necessary to apply perceptual techniques to quality measurement. Schroeder had used models to estimate the audibility of coding noise, and this technique was extended by Brandenburg to give a measure of the mean noise to masking ratio (NMR) [Brandenburg 1987]. However, this technique is ill-suited to resynthesising coders, as it relies on the concept that any difference between the two signals is noise. Karjalainen introduced a more general technique for estimating error audibility based on a comparison of audible loudness representations [Karjalainen 1985]. A similar approach was later taken up by several authors for speech quality assessment [Wang 1992, Beerends 1994, Hollier 1995] and for audio quality assessment [Beerends 1992, Paillard 1992, Colomes 1995, Thiede 1996, Sporer 1997].

At the start of the 1990s, Beerends and Stemerdink published a model, the perceptual audio quality measure (PAQM), for estimating the quality of audio coders based on a similar method [Beerends 1992]. They adapted this into a method for the measurement of speech coders known as the perceptual speech quality measure (PSQM) [Beerends 1994]. After a competition held by the Telecommunications branch of the International Telecommunication Union (ITU), PSQM was adopted as ITU-T P.861 in 1996 [ITU-T P.861] for the assessment of telephone-

bandwidth speech coders. PAQM, NMR and a number of other audio models were combined following a separate competition and subsequent collaboration in the Radiocommunications branch of the ITU to produce a new model for audio coder assessment, known as perceptual evaluation of audio quality (PEAQ), which was standardised as [ITU-R BS.1387] in 1999.

In 1992, Wang, Sekey and Gersho published a method, termed bark spectral difference (BSD), for measuring the quality of speech coders, using Sekey's perceptual loudness model as the audible loudness representation [Wang 1992]. This approach was extended by Hollier, who introduced the concept that it is necessary to consider the distribution, as well as the amount, of audible distortion [Hollier 1994, Hollier 1995].

1.3 Developments during this study

The author made a number of extensions to Hollier's model from 1997–1999, by providing an improved time alignment process, incorporating transfer function estimation and equalisation, and developing a new multi-parameter training system for model calibration [Rix 1998a, Rix 1998b, Rix 1999b, Rix 1999f, Rix 2000b]. This model was known as the perceptual analysis measurement system (PAMS), and it has been available as a commercial product since 1998. PAMS marked a key departure from previous models in that it was designed for testing not only speech coders, but also networks end-to-end, and also some acoustic testing applications.

From 1997, it became clear that PSQM [ITU-T P.861], which the ITU had standardised for testing speech codecs, was not suitable for network testing, which was the application that most telecommunications network operators were interested in [ITU-T COM12-R2]. ITU-T study group 12 therefore began a competition from 1998 to replace PSQM with a new model with much wider scope.

Five proponent companies took part in this competition: Ascom, Deutsche Telekom, Ericsson, KPN, and BT/Psytechnics. Beerends and Hekstra from KPN entered PSQM99, an improved version of PSQM that incorporated improvements similar to those that the author had already published. The author entered a version of PAMS for the Psytechnics group at BT. PSQM99 and PAMS had close overall performance, and were significantly better than the other three entries. However, both models had problems with certain types of network conditions.

To resolve these problems, the author worked with Beerends and Hekstra to combine PAMS and PSQM99 to create perceptual evaluation of speech quality (PESQ), which became ITU-T P.862 in February 2001 [Beerends 2000, Rix 2000a, ITU-T P.862, Rix 2002b, Beerends 2002]. PESQ is now in widespread use for testing telecommunications networks, in particular mobile communications and voice over packet systems.

The most recent developments have centred on acoustic testing of networks and terminals. The aim is to produce an extended version of PESQ which may be used for a range of applications where the send interface and the receive interface may each be either electrical or acoustic, and where the acoustic path may include a handset, headset or hands-free terminal connected using a head and torso simulator (HATS) [ITU-T P.58]. The author is collaborating in this work with Beerends, Berger and Goldstein, with a view to submitting a new ITU-T Recommendation in September 2003, under the working title of perceptual acoustic assessment measure (P.AAM) [Beerends 2003, Rix 2003a]. This is however beyond the scope of this thesis.

1.4 Contribution of this thesis

This thesis provides a new method to identify time-delay in the presence of non-linear and time-varying distortions. This is of similar complexity to cross-correlation but is found to be more robust to distortions such as low bit-rate coding. Two extensions of this technique to identify step delay variations, for example due to packet-based transmission, are also described.

It is shown that linear filtering, which can occur in acoustic or analogue interfaces to networks, poses a significant problem for previous perceptual models. Conventional linear estimation methods are found to be unsuitable because they are severely biased by non-linear processes such as noise, clock jitter or channel errors. More robust perceptual methods for frequency response estimation and equalisation are introduced and it is shown that these allow perceptual models to make more accurate quality predictions for conditions that include filtering.

A number of new techniques, and some standard procedures, are applied to the problem of predicting quality from a set of distortion parameters. A novel method for monotonic polynomial regression is introduced and used both for performance assessment and for normalising subjective test data to eliminate systematic non-linear variations, improving the prediction and generalisation accuracy of regression fits. Parameter selection is used in conjunction with these techniques to train a new perceptual model.

It is shown that these methods can be used to produce perceptual models that are much more accurate than before, and are applicable to a much wider range of conditions, specifically end-to-end testing of telecommunications networks that may include filtering and variable delay.

The innovations for time-delay estimation, frequency response equalisation, and model training, that are described in this thesis, have been used to develop a perceptual model, PESQ, that has been commercially successful and was standardised by the ITU for speech quality measurement of telephone networks.

1.5 Structure of this thesis

Chapter 2 presents the main background to this thesis. It sets out the scope of network conditions that may be encountered, and gives more detail on subjective testing for telecommunications and the development and application of perceptual techniques. The models PSQM, PAMS, PESQ and PEAQ are introduced. This chapter also presents performance metrics that will be used to evaluate the methods described in this thesis.

The time-delay identification problem is discussed in chapter 3. This begins with an introduction to the issues of delay variation in time and frequency. The standard signal processing techniques that can be applied for estimation of constant delay, and their limitations, are introduced. An improved method that the author developed for greater robustness to low-bit rate speech coders and time-varying distortion is presented; this uses a two-stage method based on a weighted and smoothed histogram to perform delay estimation. This is compared to a Bayesian maximum a posteriori derivation of a similar algorithm. The causes of delay variations in communications systems, and their perceptual effect, are summarised. The extension to the histogram method that the author produced for piecewise constant delay estimation, using a maximum-likelihood formulation for changepoint detection, is then described, and results are given showing the value of these techniques.

Chapter 4 focuses on the problem of linear filtering in conjunction with non-linear distortions, which had largely been ignored by previous research in perceptual speech quality assessment. This begins with an examination of the causes of filtering in telecommunications networks, with results on the effect of filtering on perceived quality and on PSQM, showing the need to incorporate transfer function equalisation into perceptual models. The main conventional (linear) frequency response estimation techniques are introduced and examples given to illustrate their weakness with noise, sample rate jitter and low bit-rate coding. Two basic perceptual transfer estimation techniques based on spectral difference and phaseless cross-spectrum are presented and applied to PESQ, along with a further improvement to address highly time-varying distortions.

Chapter 5 describes the use of multiple distortion parameters, parameter selection, multiple regression and per experiment mapping to compute the quality score. The previous technique for performance assessment of models using the logistic function is compared with a new method that the author developed, using monotonic polynomial regression. Methods for multiple linear and non-linear regression are introduced, including two new techniques for monotonic non-linear multiple regression using polynomials and the multi-layer perceptron. It is shown that regression is improved by normalisation of MOS to eliminate variations between experiments. The motivation for parameter selection for perceptual model training is presented and techniques using forward, forward-backward, and exhaustive parameter selection are

compared with McHenry's method and a commercial package. A joint training process is introduced to allow the optimum overall model to be selected. This chapter ends with a detailed set of results which show the performance of perceptual models that incorporate all of the developments described in this thesis.

Finally, chapter 6 presents the conclusions of this work and discusses potential further study in this field. Appendices list the author's relevant publications (two of which are reproduced in full), and summarise the large database of subjective tests used for these developments. Example speech recordings are provided on the accompanying CD-ROM.

CHAPTER 2

Background

2.1 Overview

This chapter sets out the main background and prior art to this thesis. In sections 2.1.1 and 2.1.2, the basic procedure of intrusive testing, and typical applications in quality assessment, are introduced. The scope of conditions that are the focus of this thesis are set out in section 2.2. Section 2.3 introduces subjective testing methods, in particular the listening quality opinion scale that is the main focus of this thesis. Section 2.4 summarises key concepts in psychoacoustics and discusses their applications in signal processing, including auditory transforms that reproduce large-scale perceptual effects, and the use of these transforms in perceptual speech and audio coders.

Three perceptual models for speech quality assessment, PSQM, PAMS and PESQ, form the main subject matter for this study, and are outlined in section 2.5. PESQ is used to produce most of the examples that are presented in the thesis. The audio quality model PEAQ is also described for comparison with the speech quality models. Finally, the methods used for measuring the performance of these models are summarised in section 2.6, and these will be used to evaluate the developments described in chapters 3–5.

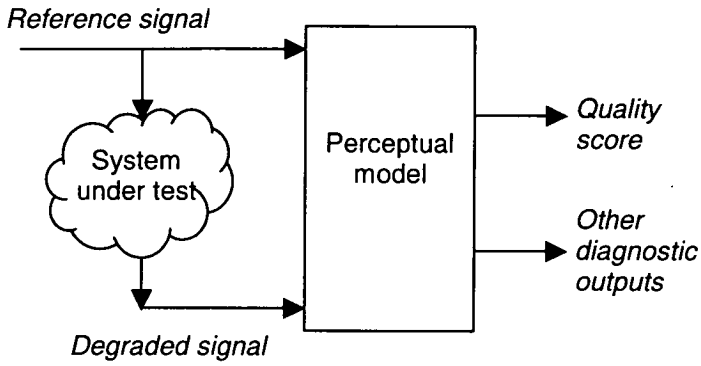
2.1.1 Intrusive testing

An intrusive measurement consists of injecting a known (reference) signal into the system under test, and recording the output degraded signal. Both signals are passed to the perceptual model, which returns a quality score, which is usually on some arbitrary continuous scale. Potentially other diagnostic outputs may be returned, for example the signal levels, spectra, frequency response, or the transforms and distortion parameters computed by the perceptual model. This basic procedure is summarised in Figure 2.1.

This thesis is mainly concerned with natural speech test signals. The communications systems that will be tested are highly optimised for speech, and components such as speech coders are known to behave in an unrepresentative way for non-speech signals such as tones, white noise, or music. A typical test signal for presentation in a subjective test consists of a pair

of sentences separated by a short pause, with total duration of about 8s and speech activity of about 50%.

Figure 2.1: Intrusive measurement using a perceptual model



2.1.2 Application overview

The following are some possible uses of a perceptual model.

Codec development and optimisation – using the perceptual model as the measure of overall performance.

Device selection – using the perceptual model, in conjunction with a large set of measurements, to provide quality information as part of the decision between technologies or manufacturers.

Network commissioning – optimising the configuration of the network for quality.

Network monitoring – making regular test calls to verify that the network meets a pre-determined quality standard, for example to raise alarms or re-route traffic if quality falls below this, or to police a service level agreement.

2.2 Scope of this thesis

Most of the techniques described here are applicable to both telecommunications and general audio transmission. However there are significant differences between the methods for subjective testing, test signals, and the underlying technologies, which mean that it has so far proven necessary to develop models targeted at either telecommunications speech quality, or general audio quality. This thesis focuses on telecommunications.

While earlier models such as PSQM were designed for assessment of speech codecs [ITU-T P.861], this thesis is intended to cover a much wider range of conditions from speech codecs to entire end-to-end communications links. Table 2.1 sets out the scope of technologies that may be included – in many possible combinations – in measurements, and for which the perceptual

model should produce accurate predictions of subjective quality. The acronyms used in Table 2.1 are defined on page xiii.

Table 2.1: Desired scope of perceptual model

Component	Options within the scope of this study	Options outside the scope of this study
Test signal	Natural recorded speech Artificial speech-like test signals Concatenated natural speech test signals Noise at talker (stationary, pseudo-stationary or time-varying) – Note 1	Music signals DTMF Network information tones
Insertion point	Digital (linear or G.711 PCM) 2-wire or 4-wire analogue Electrical connection to handset port Acoustic interface to handset/headset using HATS [ITU-T P.58]	Acoustic injection at hands-free terminals
Signal processing; network effects	Voice activity detection/discontinuous transmission (VAD/DTX) Noise reduction/suppression Automatic level control Signal level in the network Amplitude clipping in the network Quantisation Circuit noise Delay variation during a measurement e.g. due to VoIP	Acoustic echo Acoustic echo control Network echo Network echo control Conversational effect of absolute delay
Sampling rate in network	8kHz narrowband telephony	16kHz wideband telephony – Note 2
Coding	Waveform coders: G.711 logarithmic PCM (A-law, Mu-law), G.726 ADPCM CELP and related speech coders at ≥4kbit/s (e.g. G.728, G.729, GSM-EFR, GSM-AMR, EVRC) RPE-LTP coders (GSM-FR, GSM-HR)	Very low bit-rate (<4kbit/s) vocoders, including MELP and VXC – Note 3
Error types	Random and burst bit errors (waveform coders) Mobile error profiles (static, dynamic, hand-over, recorded) for mobile coders Random and burst frame erasure Random and burst packet loss	
Multiple trans-coding	Combinations of coders and error profiles	

Component	Options within the scope of this study	Options outside the scope of this study
Recording point	Digital (linear or G.711 PCM) 2-wire or 4-wire analogue Electrical connection to handset port	Stereo recordings Acoustic interface to handset/headset using HATS [ITU-T P.58] Acoustic recording from hands-free terminals
Presentation to user	Narrowband IRS/MIRS telephone handset [ITU-T P.830] at standard 79dB SPL listening level	Variation of listening level Other listening equipment e.g. loudspeakers, wideband headphones

Notes

1. For testing performance in the presence of noise, the normal method is to add noise to the speech signal prior to injection to the network. For comparison with the absolute category rating listening quality (ACR LQ) subjective testing method (section 2.3) the reference signal presented to the perceptual model should be the clean speech signal.
2. Wideband telephony is still relatively uncommon and has been studied in much less detail; however, the author has found that perceptual models can be applied to wideband telephony with minimal changes [Rix 1999e, Rix 2001a, Rix 2001b].
3. These coders may cause substantial distortion to the local timebase and/or pitch, and are normally assessed on intelligibility rather than quality.

The scope outlined in Table 2.1 is clearly very wide and it is not feasible to test all possible combinations of factors. For the P.862 competition that led to the development of PESQ, and for the P.AAM development, the ITU-T assembled a database of subjective tests where each factor in the within scope column of Table 2.1 was included, alongside other factors, in at least two subjective tests. Further details of the larger subjective test database that was used for this thesis are set out in Appendix D.

It should be noted that many of the processes listed in Table 2.1 are non-linear or time-variant, and there is such a wide range of potential behaviours that signal or system-based modelling is of very limited benefit. An advantage of the perceptual approach to quality assessment that is introduced later in this chapter is that it makes few assumptions about the processing in the system under test, and so is able to give accurate quality predictions across a wide range of conditions. However, these non-linear and time-varying effects pose significant problems for time-delay assessment and frequency response identification, for which much of the existing theory assumes that systems are linear and time-invariant. These problems are studied further in Chapter 3 and Chapter 4.

2.3 Subjective testing

A very wide range of audible distortions can be caused by the systems and processes described in the previous section. The methods for subjective testing that are introduced below

were developed to provide an overall score of the quality of a system or service from the customer's viewpoint, independent of the underlying technology used in the network. The perceptual models described in this thesis are designed to predict these opinion scores, and are therefore trained and evaluated with subjective test data.

This section provides a summary of current practice in the main telecoms research laboratories. The basic methods draw on those set out in ITU-T recommendations P.800 and P.830 [ITU-T P.800, ITU-T P.830], although these are now out of date and are being revised.

2.3.1 Listening and conversational testing

Subjective testing aims to obtain a repeatable measure of customers' perception of quality. There are two distinct classes of telephony subjective test: listening and conversational.

In listening tests, subjects hear various distorted recordings, and vote on their opinion of the quality after hearing each one. Because there is no two-way element of communication, listening tests cannot fully model the effect of listening level, talker echo, delay or handset sidetone [ITU-T G.107].

In conversational tests, pairs of subjects hold a conversation over a test network connection before voting on its quality. These measurements take account of the whole link, including handsets and sidetone, echo, level and delay impairment. Conversational tests are generally more expensive than listening tests, and a single conversational test is only able to investigate a small number of conditions.

This thesis focuses on listening models, which do not normally take account of the conversational factors: level, talker echo, delay and sidetone. These factors may be measured separately and combined with the listening quality to obtain a conversation quality measure using a model such as the E-model [ITU-T P.834, ITU-T G.107]. Conversational factors may be important in some circumstances, particularly if the network introduces significant delay and does not fully control network echo. In the remainder of this thesis, only listening quality is considered.

2.3.2 Subjective test overview

Subjective perception of quality depends on a large number of factors. In designing a subjective test it is essential to control many extraneous variables by choosing appropriate values or averaging over a typical population distribution. These methods are examined in this section.

2.3.2.1 Opinion scales

The most common technique in listening testing for telephony is known as the absolute category rating (ACR) method, using the listening quality (LQ) opinion scale set out in Table 1.1. The ACR LQ method was used for most of the subjective tests described in this thesis. In this type of test, subjects hear only the processed conditions, voting after hearing each recording. The votes given by subjects for each file are then averaged to give a file mean opinion score (MOS). The average of all votes given to all files for a given network condition is known as the condition MOS.

There are alternative test structures in use for specific applications, including the degradation category rating (DCR) and comparison category rating (CCR) methods. There is also a listening effort (LE) variant of the ACR method, asking a question based on the effort required to understand the speech sample. Because these methods use a different quality question, they will not normally give the same results as an ACR LQ test. Indeed, the author has shown that asking a different quality question may result in different conclusions being reached when comparing one type of communications technology with another [Rix 1999d].

2.3.2.2 Conditions

A typical ACR listening test allows up to 50 network conditions to be evaluated, each with speech material from four talkers. At least six of these conditions are normally given over to MNRU references [ITU-T P.810] that cover the full range of quality, and to standard network conditions such as G.711 so that quality can be compared with the existing architecture.

At the start of each test all subjects hear the same set of 6–8 preliminary conditions, covering a range of distortion types, and vote on their quality using the same procedure for voting as the main set of conditions. The votes for the preliminaries are discarded; they serve as an anchor to ensure that all subjects start the test with the same idea of what the range and types of distortion will be.

2.3.2.3 Other factors

A test aims to obtain a measure of the subjective quality of a number of network conditions. However there are usually many other variables, which must be controlled through the test design. These include the following dependencies.

- Talker gender, dialect/accent and other characteristics
- Source/sentence material
- Presentation order
- Balance of conditions in the test

- Individual subjects' preferences
- Quantisation and random errors in the voting process

One major factor that cannot be controlled in a single subjective test is dependence on the language and culture of the subjects. Subjects are normally native speakers of the language used in a test, and the interpretation of the opinion scale and the experience of existing telephone networks vary from country to country. This, combined with the effects of balance and individual preferences, are the main causes of the variation between subjective tests that is described in sections 2.6.4 and 5.3, and which poses a significant challenge for perceptual model training.

2.3.2.4 Processing of speech material

Although the methods used to process material for a subjective test are beyond the scope of this thesis, the following are examples of the processing stages for simulating a telephone network condition [ITU-T P.830].

1. Record original speech material using high-quality microphone in quiet conditions.
2. Apply send filtering (e.g. MIRS) and level alignment (e.g. to –26dBov [ITU-T P.830]).
3. Add environmental noise at appropriate level if required.
4. Downsample to 8kHz.
5. Apply coder.
6. Channel error insertion.
7. Apply decoder.
8. If multiple transcodings are simulated, a filter and an arbitrary delay may be inserted to make the transcodings asynchronous, then the coder/error/decoder stages are repeated.
9. Upsample to 16kHz for presentation in subjective test, checking for clipping.
10. Verify that active speech level lies in desired range.

2.3.2.5 Presentation to the subjects

Subjects are briefed according to a standard procedure, describing the process and the opinion scale [ITU-T P.800]. The recorded material in an ACR LQ test is presented to subjects over a telephone handset with a standard MIRS receive frequency response [ITU-T P.830], in an acoustically isolated room with a controlled noise level. After the preliminaries, the presentation order is randomised for each subject, to minimise the effect of order dependence.

2.3.2.6 Analysis of results

The average of all of the subjects' votes for a given network condition is known as the condition mean opinion score (MOS). In some cases it is also useful to consider the average for each talker, or separately for male and female talkers, to assess dependence on accent or gender. A number of statistical techniques may be deployed to analyse the significance or otherwise of the differences encountered, including analysis of variance and pair-wise t-tests [Duckworth 1968].

2.3.3 Limitations of subjective testing

Subjective tests provide a controlled and reproducible method to assess customer opinion of the quality of audio communications systems. However, they are far from perfect. Repeatability of absolute scores is one problem: in P.800 tests, naïve listeners are generally preferred, but the lack of listener training means that large variations in opinion occur between different subjects, tests, or laboratories. This is discussed further in section 2.6.4, and means that it is difficult to compare results directly from one test to another.

To obtain controlled conditions requires the use of specialised, acoustically isolated listening rooms and calibrated presentation equipment, which only a small number of laboratories world-wide are able to provide. Each subject typically spends one or two hours' listening time for an ACR LQ test, making it expensive and slow to use large numbers of subjects. The confidence intervals on MOS are roughly inversely proportional to the square root of the number of subjects. Testing therefore requires a trade-off between cost and statistical accuracy.

Together, these problems of speed and cost mean that subjective testing is often only commercially viable for large decisions, such as the approval of new standard codec or a carrier-scale purchase of network equipment, and is impractical for day-to-day applications in research and development or network monitoring. This provides the motivation for the development of computational models that predict subjective quality, which can be used in place of subjective tests for many applications.

2.4 Perceptual signal processing

This section discusses the processing of signals in the human auditory system, signal processing models of these processes, and their use in perceptual audio and speech coding.

2.4.1 Psychoacoustics

Research in the experimental psychology of hearing has been conducted for more than 100 years and has built detailed models of many of its properties. Whilst it is not the purpose of this section to provide a comprehensive review of this field, it is valuable to understand those

properties of hearing that are relevant to the perception of quality, in particular the audibility and perceived loudness of distortions. More detailed descriptions of the human auditory system can be found in [Moore 1997a] and a review of key results in speech perception is given in [Handel 1989].

2.4.1.1 Absolute threshold of hearing

There is a minimum sound pressure level below which it is impossible to discriminate sounds. The absolute threshold of hearing is defined as the average level at which subjects can correctly identify the presence of a pure tone, typically of duration on the order of 1s, in a silent environment. At 1kHz, for a subject in their early 20s with unimpaired hearing, listening binaurally, this level is taken to be $2 \cdot 10^{-5}$ Pa measured in the free field. It should be noted that the threshold can vary by as much as ±20dB between different people, so it is usual for results to be given as averages across a population of listeners. The test method and specified detection rate also have an effect on the measured threshold.

The threshold of hearing also varies with frequency. At 100Hz it is elevated by about 20dB; at 30Hz, by about 40dB. The threshold, particularly at higher frequencies, also rises with age and with hearing damage [Moore 1997a].

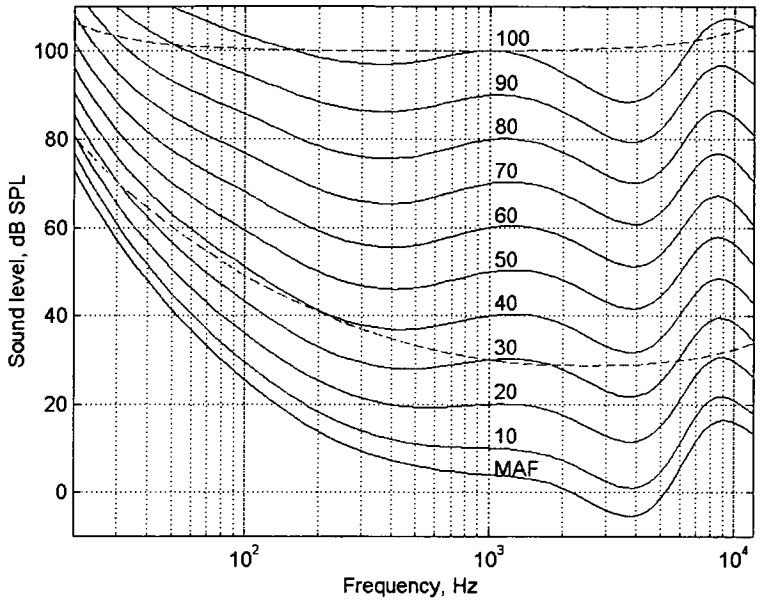
2.4.1.2 Loudness perception against frequency

The unimpaired human auditory system has a very wide dynamic range – up to 120dB from the threshold of hearing to the threshold of pain. Since this range is so wide, it is convenient to measure sound levels on a logarithmic scale. The dB SPL (sound pressure level) is defined in decibels relative to the absolute threshold at 1kHz, so 0dB SPL corresponds to $2 \cdot 10^{-5}$ Pa.

Just as the threshold varies with frequency, so does perception of loudness. This is commonly measured by presenting subjects with tones alternating between two frequencies, and asking them to adjust the loudness of one until it is perceived to be as loud as the other. The phon equal loudness scale is defined as the dB SPL level of a tone at 1kHz that is of equivalent loudness to the sound in question. For quiet sounds, the absolute threshold is the dominant effect, so a sound just above threshold is perceived to be of roughly equal loudness independent of frequency. Thus at 100Hz, a sound of 10 phon is at a level of about 30dB SPL, mirroring the elevation in threshold of 20dB at this frequency.

This does not hold at all frequencies, however. In particular, there is compression in the phon scale at low frequencies. The recruitment effect, which is thought to be due to the non-linearity of the physiological and cognitive processes, means that a sound at 100Hz at 100 phon is at about 103 dB SPL – only 3dB above the equivalent level at 1kHz. The phon scale and the minimum audible field (MAF) – another term for the threshold of hearing – are shown in Figure 2.2, which is based on [ISO 226].

Figure 2.2: Phon equal loudness scale



For the measurement of perceived sound intensity, the effect is modelled by filters with a frequency response that loosely follow the equal loudness contours at given levels: the A-weight filter at 30 phon, the little-used B-weight filter at 70 phon, and the C-weight filter at 100 phon [Moore 1997a, IEC 60651]. The A-weight and C-weight filter shapes are shown by the dashed lines in Figure 2.2.

2.4.1.3 Absolute loudness perception

Although dB SPL is a convenient engineering measure, a logarithmic scale is not a good model of the perception of loudness. The sone loudness scale is derived by asking subjects to alter the level of a tone so that it is, say, twice as loud as another tone at the same frequency, with the two presented alternately. An approximation found by Stevens is that sone loudness is proportional to intensity raised to the power 0.3 [Stevens 1972]. Since $10^{0.3} \approx 2$, 10dB corresponds to a doubling in the perceived loudness of a sound.

For complex sounds with components at many frequencies, it has been found that a better approximation of the total perceived loudness may be obtained by summing the sone level in a number of frequency bands, rather than by measuring dB SPL using the appropriate weighting filter and converting the result into sone [Stevens 1972, Sekey 1984].

Variants of the sone scale, and the associated frequency-domain processing to compute loudness of complex sounds, have been published by several authors, including Moore [Moore 1997b], Sekey [Sekey 1984] and Zwicker [Zwicker 1990].

2.4.1.4 Loudness discrimination

With alternate presentation of the same sound at different levels, it has been found that subjects can discriminate level variations on the order of 1dB at most levels above threshold, which is known as Weber's law. The ability to detect amplitude modulation is also at a similar level [Moore 1997a]. However, this works only with alternate or continuous presentation. In absolute listening (without a reference) to sounds where the level may vary naturally – in particular, speech – the threshold to detect variation in loudness can be 10dB or higher.

2.4.1.5 Simultaneous (frequency) masking

If two tones of similar frequency (but without a common fundamental frequency) are presented, the louder tone may make the quieter tone completely inaudible – i.e. masked. For very close frequencies, the threshold is between 8–13dB, depending on whether tones or narrowband noise are used. The effect becomes weaker as the frequency spacing rises; this decay is asymmetric, with the dominant effect being that components of higher frequency tend to be masked. The effective threshold of hearing in the presence of a number of signal components is termed the masked threshold, and it can be estimated using models described in the next sections [Moore 1997a].

2.4.1.6 Frequency resolution

Above about 100Hz, the perceived spacing of two tones is approximately determined by the ratio of their frequencies, rather than their linear difference. This leads to the definition of the musical scale in terms of logarithmic frequency – for example, an octave is a doubling of frequency. This effect is also reproduced in simultaneous masking, where the masked thresholds for tones of different centre frequencies appear of similar shape when plotted on a logarithmic frequency scale. Models of simultaneous masking and frequency perception have led to the development of perceptual frequency scales such as the bark (critical band) and equivalent rectangular bandwidth (ERB) scales, taking account of the logarithmic relationship above 100–300Hz and roughly constant resolution below this [Zwicker 1990, Moore 1997a].

2.4.1.7 Temporal masking

In a similar way, masking can apply to two sounds separated in time. The strongest effect, forward masking, is that a loud sound raises the threshold for sounds occurring after it. This may last up to 150ms [ITU-R BS.1387]. The decay rate varies with frequency, absolute level, and the duration of the masker [Moore 1997a]. The converse effect, backward masking, is weaker, decaying in about 5ms [ITU-R BS.1387]. According to this effect, a sound may be masked by a louder sound following it within a few milliseconds. Backward masking is a particular problem with early perceptual audio coders, that used a fixed frame size, at sharp

signal onsets. This is because the DCT causes quantisation noise to be spread throughout the frame duration, which is typically 20–40ms, leading to audible distortion known as pre-echo.

2.4.1.8 Perceptual streaming

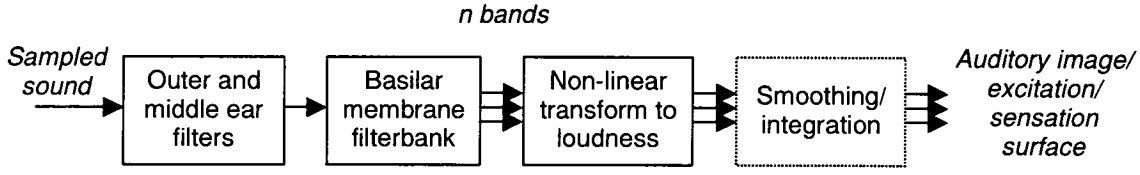
A higher-level cognitive effect known as perceptual streaming is also relevant to quality measurement. The brain appears to group sounds that come from a single source, such as a talker, into one continuous stream. Sound events that come from other sources – for example, with different frequency content, amplitude envelope, or spatial location – are perceived as distinct streams. Often it is difficult to resolve the exact temporal relationship between the two streams. The effect does not happen for components that are deleted in time or frequency, which may be interpolated by the brain from the other content [Handel 1989, Bregman 1990, Moore 1997a]. For perceptual quality modelling, this difference between additive and deletive distortions was termed the asymmetry effect [Beerends 1994].

2.4.2 Auditory transforms

The properties described in the previous section may be approximated digitally. This is important because it means that the properties of hearing can be used to optimise speech or audio coders, or the perceptual quality assessment models that are the focus of this thesis.

The generic structure of an auditory transform – analogous to a spectrogram – that is used by most authors is shown in Figure 2.3. The output of this transform is a representation, for time t and frequency f, of the partial loudness of the signal, or the intensity of neural firings for the channel of centre frequency f at time t. With loudness models, the instantaneous loudness is obtained by integrating over frequency. Surveys of auditory transforms are given in [Cooke 1993, Cooke 2001].

Figure 2.3: Generic auditory transform



Auditory transforms such as this are used to predict the masked threshold, loudness, and pitch perception [Patterson 1992]. They also form the input to models of higher-level cognitive processing of acoustic events such as [Cooke 1993], and are central to the comparison method for perceptual quality modelling that is the focus of this thesis.

The outer and middle ear filters model the linear filtering of sounds from free field due to the acoustic shape of the head and upper body, the pinna, and ear canal. This filtering accounts for

much of the shape of the lower threshold of hearing. A bank of filters is used to model the time-frequency analysis that is performed by the hair cells along the basilar membrane in the inner ear. This is often approximated by a linear filterbank, or by transformation from the STFT, although there is evidence that the physiological filters vary with level [Brandenburg 1987, Glasberg 1990]. The outputs of these filters are usually chosen to be equally spaced on a perceptual frequency scale such as the bark or ERB scales [Moore 1997a].

The final stages of the transform differ depending on whether it is used to model excitation, loudness, masked threshold, or a more detailed characteristic such as the neural response or binaural processing. Loudness transforms may use a non-linearity that maps directly to a sone scale [Zwicker 1990] or from phon to sone [Stevens 1972, Sekey 1984]. Masked threshold models may include an empirical estimate of the tonality of each component, to allow for the different masking effects of noise and tones that is thought to be due to phase synchrony in the auditory system [Moore 1997a]. For neural models, the non-linearity may be used to represent the phase sensitivity of the hair cells as well as the threshold and loudness response. Adaptive gain control has also been used here to account for cross-band and temporal masking properties. In binaural models, cross-correlation between the channels for the two ears is used to obtain spatial cues such as inter-aural time difference [Martin 1995, Moore 1995, Cooke 2001].

2.4.3 Perceptual coding

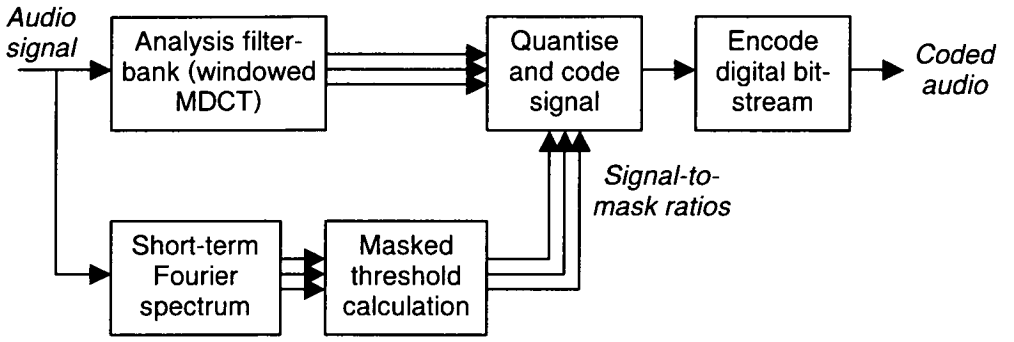
2.4.3.1 Perceptual audio coding

A powerful application of auditory modelling is its use in perceptual audio coders such as the MPEG family [Brandenburg 1996]. High-quality uncompressed audio requires very high bit-rate to transmit – 1.4Mbit/s for CD-quality audio. Lossless compression schemes such as Huffman coders are not able to reduce this by more than a factor of two or so for general signals. However, the use of a model which estimates the masked threshold allows efficient lossy compression to be performed.

This is illustrated by the simplified MPEG audio coder structure that is shown in Figure 2.4 [Rix 1999a]. The coder works on overlapping windowed frames of input signal. The auditory transform is used to estimate the signal-to-mask ratio at each frequency. This determines the allocation of bits, with the aim that the quantisation noise always lies below the masked threshold. The decoder, which does not include a perceptual model, decodes the bit-stream and runs the inverse of the analysis filterbank to reconstruct the decoded audio. In conjunction with other components, such as joint stereo coding, long-term prediction, perceptual noise substitution, and window-length switching to ensure that time-domain aliasing remains

temporally masked, recent variants of this coder are able to offer 24:1 compression, at 64kbit/s for stereo audio, with little audible degradation to the signal quality.

Figure 2.4: Simplified MPEG audio coder



2.4.3.2 Perceptual speech coding

For the same reasons – reduced cost of transmission and storage – it is often desirable to compress speech signals. The development of low bit-rate speech coders pre-dates that of perceptual audio coders; the basic concept of masked noise loudness used in the MPEG audio coders was first proposed for speech coding [Schroeder 1979].

The CELP structure that is in most common use today also processes frame-by-frame, though often without overlap. In this case an iterative search is performed for each frame. The coding error for each codebook entry is found by synthesising the decoded signal, and a simple perceptual masking model is used to compute a distortion metric, leading to the term for this structure, analysis by synthesis. This allows the optimum perceived quality to be achieved for the given codebook and coder structure in coders such as G.728, G.729, GSM-EFR and GSM-AMR. These provide good quality for voice telephony at bit-rates on the order of 5–16kbit/s.

2.5 Perceptual models for intrusive quality assessment

This section describes the main background to this thesis, the development of perceptual models for speech quality assessment. Beerends' PSQM and Hollier's PAMS, which both pre-date this research, are introduced. Developments to PAMS during this study, and the new model PESQ which was produced from a collaborative integration of PAMS and PSQM, are summarised. The audio quality model PEAQ, which was published about a year after this research began, is also described.

2.5.1 History

Two contrasting approaches for intrusive perceptual quality assessment emerged in the 1980s. Schroeder et al proposed a comparison of the error signal to the masked threshold [Schroeder 1979] for use in a perceptual speech coder. Karjalainen proposed instead to compare auditory transforms of the reference and degraded signals to identify errors [Karjalainen 1985]. The first major quality assessment model that followed the philosophy of Schroeder was noise-to-mask ratio (NMR), which considers the time-domain difference between the two signals to be noise, and essentially determines whether this is audible by comparing the short-term noise spectrum to a masking model [Brandenburg 1987].

In NMR, the error signal (the difference between the reference and degraded signals) is calculated for each frame. This is transformed to the frequency domain and compared with a worst-case masked threshold computed from the reference signal. The ratio, NMR, is averaged over frequency and the duration of the measurement [Brandenburg 1987] to give an estimate of the audibility of distortions. NMR was one of the models incorporated into PEAQ, which is described below.

NMR has some fundamental disadvantages. While it may provide a reasonable estimate of distortions close to or below the masked threshold, it is not designed to take account of more clearly audible distortions where the amount and annoyance of distortion is the main factor. In particular, it does not include an asymmetry effect. Because NMR essentially processes the time-domain error signal, it is not robust to effects such as phase shifting, amplitude modulation, linear filtering or to resynthesising (non-waveform) coders, which create a similar sound that may nevertheless differ substantially in the time domain [Robert 1999]. In this case a large error is measured even though the difference may be perceived as minimal. Finally, the structure of NMR closely follows that of the coders that it is used to assess. This means that it is ill-equipped to analyse artefacts that are poorly modelled by the coders, such as time-domain aliasing.

An alternative approach, that of comparing internal representations – i.e. auditory transforms – of the reference and degraded signals, was first introduced by Karjalainen [Karjalainen 1985]. Unlike Schroeder et al, this was proposed for calculation of an overall quality score, the auditory spectrum difference (ASD). In this method, the error is obtained from a comparison of the loudness of the two signals derived using a filterbank. Masking is reproduced by the compressive nature of the loudness transform, combined with the frequency response of the perceptual filters and a non-linear time-domain spreading. Re-synthesis is less of a problem as the gross perceived loudness, rather than the time-domain waveform, is compared. Because it is possible to use a more complex transform than in speech or audio coders, effects such as gain variation and temporal masking may be modelled more accurately, and it is straightforward

to model the asymmetry effect. For these reasons the method of comparison of transforms was taken up by several authors and has been much more successful.

Sekey's loudness model was applied to create the bark spectral distortion (BSD) measure of overall quality by Wang [Sekey 1984, Wang 1992]. In some ways this was further developed than ASD, but lacked several of the latter's features including temporal masking. BSD was further extended by Hollier to create the perceptual analysis measurement system (PAMS) [Hollier 1994, Hollier 1995], which is outlined below and formed the basis for the research described in this thesis. Beerends developed an alternative family of models drawing on Zwicker's loudness model [Zwicker 1990]. The perceptual audio quality measure (PAQM) [Beerends 1992] was later adapted – including the removal of frequency-domain smearing – to give the perceptual speech quality measure (PSQM) [Beerends 1994, ITU-T P.861], which is also described below. The scaling and asymmetry processing in PSQM was later extended by Beerends [Beerends 1997], and this was used for the auditory transform in the perceptual evaluation of speech quality (PESQ) model, which the author co-developed with Beerends and Hekstra [Beerends 2002, Rix 2002b, ITU-T P.862], and which is used for many of the examples described in this thesis.

Other authors have also extended BSD. Yang introduced a masking model to ignore errors below the masked threshold [Yang 1997]. Novorita experimented with window size and forward and backward masking, concluding that incorporating both effects using an exponential decay function improved accuracy [Novorita 1999]. An alternative error computation using phaseless coherence in the bark spectrum, which automatically eliminates the effect of frequency response, was proposed by Park [Park 2000], drawing on concepts that the author had already proposed for PAMS [Rix 1999f] and that are discussed in Chapter 4.

It is interesting to note that it was only during the late 1990s that perceptual models were generally accepted as more accurate predictors of MOS than simple objective measures. BSD and PSQM were primarily compared against non-perceptual models including signal-to-noise ratio (SNR), segmental SNR, LPC log likelihood ratio, mean cepstral distance, information index, and pattern recognition, which had been described by many authors [Quackenbush 1988, Kitawaki 1988, Kubichek 1989, Lalou 1990, Kubichek 1991, Kubichek 1993, Takahashi 1996]. The most recent of this family, the spectrogram-based measuring normalising blocks (MNB) model, was briefly standardised as an appendix to P.861 from 1998 until both PSQM and MNB were replaced by PESQ [ITU-T P.861, Voran 1999a, ITU-T P.862]. At the time of writing, perceptual models are only starting to be used for image and video quality measurement, although the benefits of perceptual methods in these applications are similar to those for speech and audio.

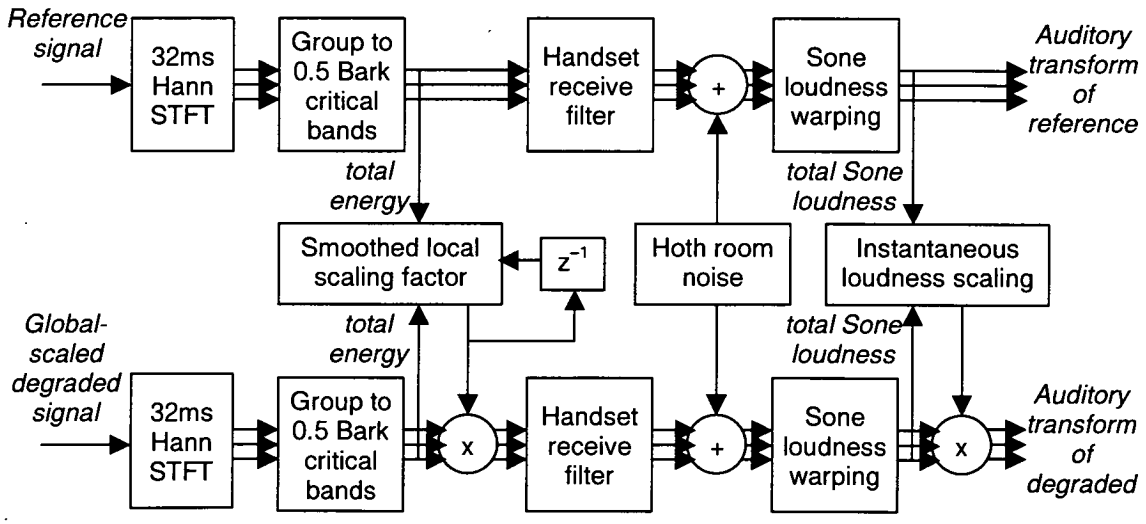
2.5.2 Perceptual speech quality measure (PSQM)

PSQM was proposed by Beerends and Stemerdink in 1994 for the assessment of telephone-band speech coders [Beerends 1994], as an extension of their audio quality model PAQM [Beerends 1992]. It became the first perceptual quality model to be standardised, after a competition run by the ITU-T from 1994–96 [ITU-T P.861]. This section gives an summary of the structure, scope and limitations of PSQM, and is based on [ITU-T P.861].

At the core of PSQM is the pair of auditory transforms shown in Figure 2.5. The basic transform is applied to both reference and degraded signals. It consists of Hann windowed STFT on 32ms 50% overlapping frames. The power is then grouped, without smearing or interpolation, into critical bands of about 0.5 bark, equally spaced on a bark perceptual frequency scale, with 42 bands at 8kHz sampling rate. A filter is applied in the bark spectral domain to model an IRS receive telephone handset, and noise is added to simulate 45 dB SPL(A) Hoth noise in the listening room. The power in each band is then warped to a compressed sone loudness scale. This is based on that of Zwicker [Zwicker 1990], but has an unusually low exponent γ=0.001.

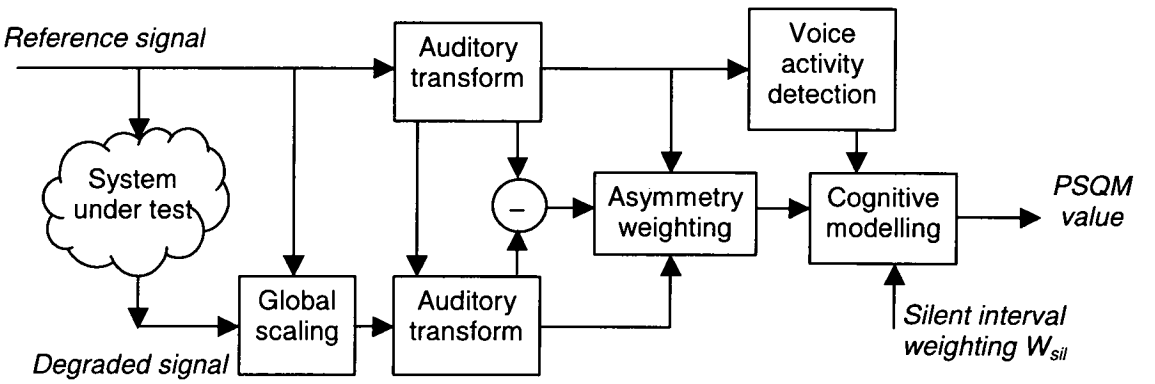
Some additional processing is also applied to adapt the degraded signal to the reference. Slow gain variations in the system under test are eliminated by applying a scaling factor, smoothed with a first-order filter, on the degraded signal. This partially compensates for time-varying gain: slow variations are almost completely eliminated, but more audible rapid variations in gain are largely uncorrected. After the loudness warping, each frame of the degraded signal is scaled to be of equal loudness to the reference; however, because of the non-linearity in the loudness scaling, this does not cancel out the partial equalisation that is applied in the power domain.

Figure 2.5: PSQM auditory transforms



The auditory transforms are applied in PSQM as shown in Figure 2.6. The degraded signal is pre-processed to scale it to the same RMS level as the reference signal. After the auditory transform, the noise disturbance is calculated as the absolute difference in sone loudness, with a deadzone of 0.01 sone. The disturbance is summed over frequency using an asymmetry weighting calculated from the ratio of the signal powers after Hoth noise addition, raised to the power 0.2 (see equation (5-3) in section 5.2.2). This amplifies loud additive distortions, modelling the asymmetry effect. Further explanation of this process is given in Appendix C. The weighted disturbance is linearly averaged in time, separately for speech and noise periods using a simple VAD. These are combined with a silent interval weighting factor W_{sil} to give a single distortion measure known as the PSQM value, which is in the range [0, 6.5]; zero corresponds to no distortion. No final value for W_{sil} was agreed in P.861: it was suggested that this should be adapted for each subjective test to account for differences in the weight that subjects gave to distortions in the silent periods, but in practice the provisional value of 0.2 was a reasonable compromise [Rix 1998e] and has been used in most implementations.

Figure 2.6: PSQM model structure



The standardisation of PSQM as P.861 [ITU-T P.861] focused on its application for predicting the quality of speech coders. However, it was found that the main application was for testing telephone networks, which led to problems with the accuracy of PSQM. Firstly, time alignment and level alignment methods were not defined, although these components must give the correct results for the model to behave as designed. Secondly, with no transfer function equalisation process, PSQM overestimates the distortion for systems that include linear filtering, such as 2-wire telephone networks. Thirdly, the model often gives highly inaccurate results if noise, channel errors, or delay variations are introduced by the system under test. These limitations formed the motivation for the study described in this thesis.

2.5.3 Perceptual analysis measurement system (PAMS)

BSD [Wang 1992] was extended by Hollier to compute a wider range of distortion parameters [Hollier 1994]. With the provision of time-alignment based on cross-correlation of the two signals, and level alignment to a fixed active speech level, this became known as PAMS and formed the test bed for the work described here. The auditory transform used at the core of PAMS is shown in Figure 2.7. The overall structure of PAMS, as initially developed by Hollier up to 1995, is shown in Figure 2.8.

Figure 2.7: PAMS auditory transform

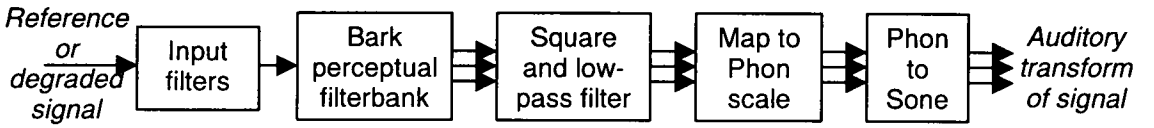
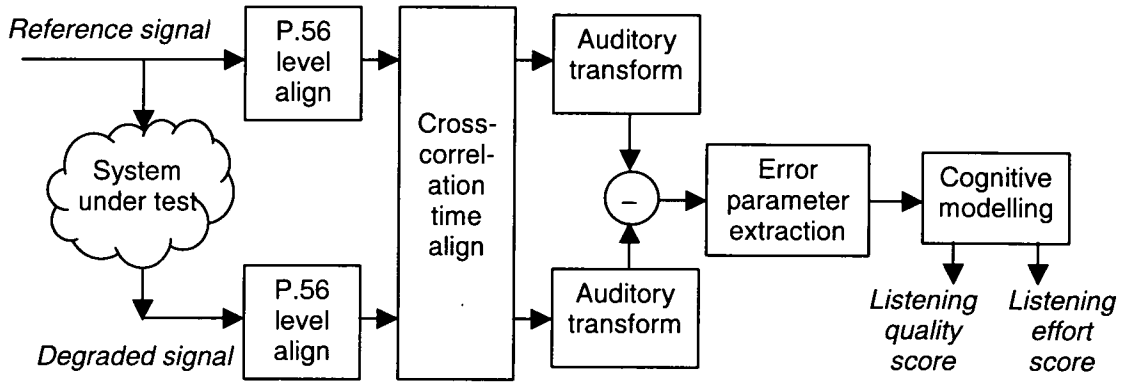


Figure 2.8: Initial structure of PAMS



In Hollier's model, processing begins by aligning both signals to the same, standard –26dBov active speech level using the P.56 method [ITU-T P.56]. The delay between the two signals is estimated by cross-correlation, and is eliminated in further processing. The auditory transform starts with a set of linear filters which model the MIRS receive frequency response of a standard telephone handset [ITU-T P.830], the leakage in the acoustic coupling between the earpiece and ear, and the frequency response of the middle ear. This is followed by a bank of deeply-overlapping filters of approx. –3dB bandwidth 1 bark, equally spaced at 1 bark intervals. Frequency masking is reproduced by the filter shapes, which decay at –10dB/bark below the centre frequency, –25dB/bark above this.

The outputs of the filterbank are squared and filtered through a rectangular low-pass filter of duration 4 cycles at the centre frequency, or 8ms, whichever is the greater, and downsampled to 4ms. The resultant power is mapped to a phon scale and then to a sone scale. The filterbank shape and spacing follows [Sekey 1984], and the mappings to phon and sone scales follow

[ISO 226, Stevens 1972]. For comparison, BSD implements the filterbank with a windowed DFT on 10ms 50% overlapping frames, rather than as individual filters, and uses a simpler approximation for the phon mapping.

PAMS further differs from BSD in the computation of distortion parameters. BSD calculates the total squared sone difference over frequency for each frame, and averages this over all active speech periods. Silent intervals are ignored in BSD, and [Wang 1992] also suggests ignoring unvoiced sections to improve accuracy. This is very unsatisfactory for an overall quality assessment measure. In PAMS, Hollier and Gray introduced the computation of several different parameters, error activity (mean absolute error), positive error activity (mean error with negative components set to zero), negative error activity, and error entropy. These were combined using empirical methods or standard linear regression techniques to obtain a more accurate quality measure. Two quality scores in the range [1, 5] are computed, one on the ACR listening quality opinion scale shown in Table 1.1, known as Y_{LQ}, the other on the listening effort opinion scale, Y_{LE}.

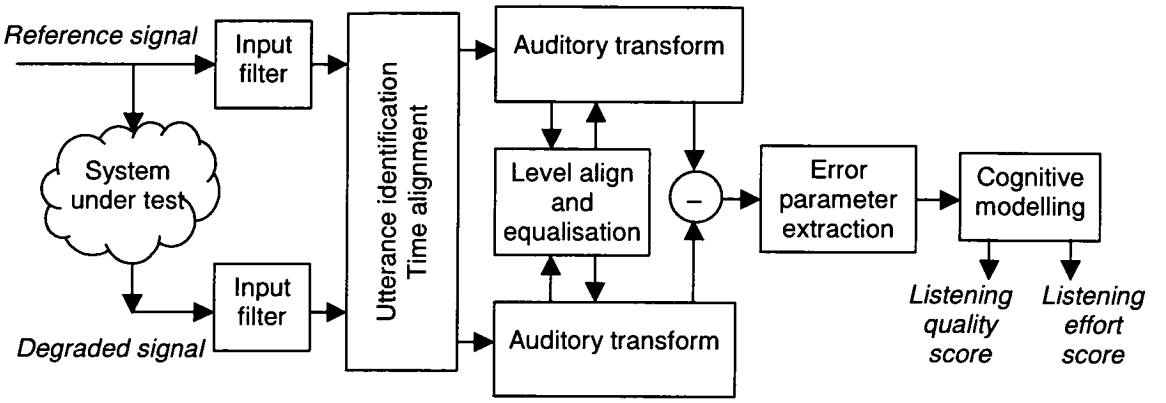
While PAMS, at the start of this project, was further developed than PSQM, it shared many of its limitations. The cross-correlation time alignment proved to be slow and gave large errors with some low bit-rate coders, particularly in error conditions. The lack of a process for transfer function equalisation meant that PAMS also over-estimated the effect of linear filtering. Finally, accuracy became poor with noise or delay variations in the system under test.

The author made a number of improvements to PAMS as part of the work described in this thesis. The following are the most significant points.

- The time alignment process was replaced by the piecewise constant delay identification method that is described in Chapter 3, allowing PAMS to be used for measurement of systems that exhibit variable delay.
- Transfer function estimation and equalisation were included as part of the auditory transform (prior to the phon mapping) to improve accuracy with systems that include linear filtering; this is discussed further in Chapter 4.
- Parameter selection was used to develop a more accurate cognitive model, using the techniques described in Chapter 5.

The overall structure of PAMS including these extensions is shown in Figure 2.9. Further details on PAMS are presented in the paper that is reproduced in Appendix B.

Figure 2.9: PAMS structure as developed by the author



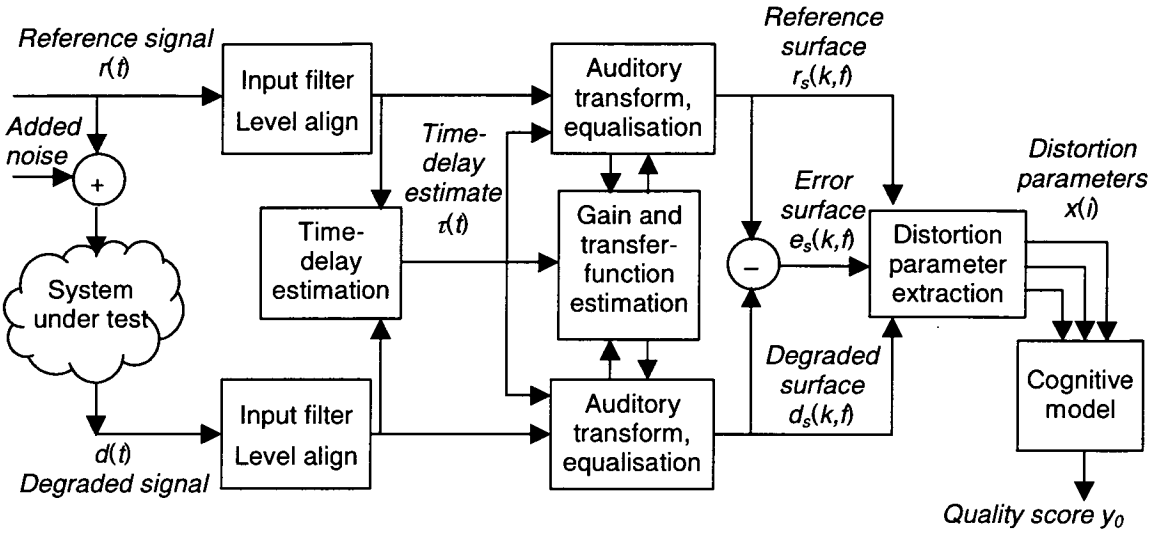
2.5.4 Perceptual evaluation of speech quality (PESQ)

The limitations of PSQM prompted the ITU-T to run a competition to standardise a new perceptual model, with the scope described in section 2.2 [ITU-T COM12-R2]. Five models were entered. The author submitted PAMS. Beerends and Hekstra, from KPN, entered a model known as PSQM99, an improved version of PSQM, that incorporated many of the enhancements described for PAMS. Berger, from Deutsche Telekom, submitted TOSQA [Berger 1997]. Submissions were also made by Juric, from Ascom, and Karlsson from Ericsson.

The overall performance of PSQM99 and PAMS was much higher than any of the other models, but neither met all of the ITU's performance requirements for unknown data [Klaus 2000]. PSQM99 had poor correlation for subjective tests including VoIP or significant filtering; PAMS gave incorrect results with gain variations or noise switching during silent periods. To produce a single model that would be acceptable to the ITU, the author collaborated with Beerends and Hekstra in March–May 2000 to combine PSQM99 and PAMS to produce PESQ, which was standardised as P.862 the following year following independent evaluation by a number of telecoms research laboratories [Beerends 2000, Rix 2000a, Rix 2000c, ITU-T P.862, Rix 2002b, Beerends 2002]. A detailed description of PESQ is given in [Rix 2000c], which is reproduced in Appendix C.

The best performance was found to be achieved by combining the variable-delay time alignment process from PAMS with the auditory transform and cognitive model of PSQM99. The combined model was optimised using the new subjective tests that had been conducted for the competition. The simplified structure of the combined model is shown in Figure 2.10.

Figure 2.10: PESQ model structure



The processing in PESQ begins with input filtering and level alignment to a standard listening level corresponding to about 79dB SPL. The utterance time-delay estimation method described in Chapter 3, including utterance splitting, is used to identify the delay and align the auditory transform frames between the two signals. The auditory transform has similar basic structure to Figure 2.5, but without handset filtering or Hoth noise insertion. The Zwicker exponent γ, in contrast to PSQM, is set to 0.23, which is a typical value from the literature [Zwicker 1990]; this is adjusted at low frequencies for the recruitment effect. Further changes to the auditory transform include frequency response equalisation, which is performed using a modified bark spectrum difference (described further in section 4.4) prior to local scaling for equalisation of time-varying gain.

PESQ differs significantly from PSQM in the final stages of the model. The deadzone is made proportional to the signal loudness in the given time-frequency cell, which more closely follows Weber's law. Two separate distortion parameters are computed: a symmetric disturbance measure based on the RMS distortion in each frame, and a mean asymmetric disturbance measuring only loud additive errors, to model the asymmetry effect. A further innovation by Beerends is a two-stage temporal averaging process, where each disturbance measure is first averaged using an L_p-norm with p=6 on 50% overlapping rectangular windows of about 300ms duration, and the results are averaged with p=2 (RMS) over the duration of the signals (Appendix C). The high p for short-term averaging gives most weight to the largest individual distortions, giving a simple model of perceptual streaming. These two distortion parameters are combined in a linear output stage, to give a PESQ score on an scale of [–0.5, 4.5], where 4.5 corresponds to zero distortion. PESQ also includes a process to realign critical cases, which is described in section 3.5.6.

PESQ has proven to be remarkably robust; only a relatively small number of problem cases have emerged in the three years since it was completed. Many of these relate to extreme cases that were not planned for during the model's development, and lead to unexpected results: where the reference and degraded files differ in length, where there is more than 25% drop in speech activity, or where completely unrelated reference and degraded files are compared. However, some cases that are more likely to be encountered in real use have been found where the model may give inaccurate scores. The local gain compensation process appears to interact with muting distortion, which typically occurs in the presence of front-end clipping due to VAD or as a consequence of unconcealed packet loss, leading to quality scores that may be too low in the presence of VAD but too high with packet loss. Other issues include processing of delay variations during speech, which is discussed in section 3.5.7, and extreme cases of linear filtering, which is considered in section 4.4.3.

At the time of writing (May 2003), the author is currently involved in a new collaboration with Beerends, Berger and Goldstein to develop an enhanced version of PESQ with an extended scope that includes the acoustic path and terminal equipment at either the transmit or receive ends, or both [Rix 2003a]. The working name of this project is P.AAM (acoustic assessment model), and it is expected that the model will be standardised as a new ITU-T recommendation in September 2003. While the new developments in P.AAM are beyond the scope of this thesis, a working version of P.AAM by the author was chosen to extract the distortion parameters that are used to demonstrate model training in Chapter 5.

2.5.5 Perceptual evaluation of audio quality (PEAQ)

From 1994–96 the ITU-R ran a competition for the selection of a model for audio quality evaluation. This model was required to predict subjective difference grade (SDG), obtained using subjective tests conducted according to BS.1116 [ITU-R BS.1116]. Unlike the concurrent speech quality competition, the six entrants were all based on perceptual criteria. NMR [Brandenburg 1987] was the only masked-error model; the remaining five use the method of comparison of internal representations. Of these, PAQM [Beerends 1994], PERCEVAL [Paillard 1992] and Toolbox [ITU-R BS.1387] are based on the DFT, while DIX [Thiede 1996] and POM [Colomes 1995] use filter banks. No single model showed substantially better performance or met all of the competition's requirements, so a collaboration phase followed from 1996–98. A seventh model, OASE [Sporer 1997], was incorporated at this stage. The resulting standard, known as PEAQ was approved by ITU-R in 1999 [ITU-R BS.1387]. The description here, based on [ITU-R BS.1387], is provided because some of the concepts that it incorporates are also of value for perceptual speech quality assessment, which is the main focus of this thesis.

PEAQ consists of two models, both operating in stereo at 48kHz sampling rate. The Basic version uses a 42ms Hann windowed DFT, with 50% overlap, for the auditory transform, and is

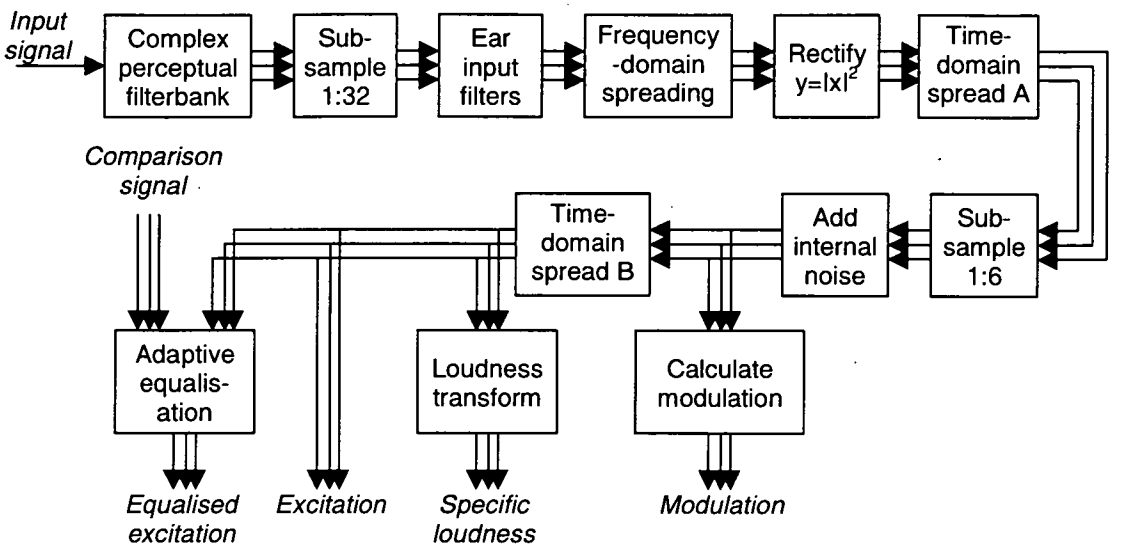
designed for real-time implementation. The Advanced version uses results from a slightly modified form of the Basic version, but also uses a filterbank auditory transform based closely on that of DIX [Thiede 1996]. The Advanced version is therefore somewhat slower than the Basic version, but is more accurate. With the exception of the reduced temporal resolution of the DFT-based transform, most of the processes and the calibration of the two models are very similar. The following considers the filterbank auditory transform.

Prior to processing, the input signals must have been time-aligned. No algorithm for this is provided. With high-quality audio using waveform coders, time alignment is easier than for telecommunications. The main problem is slow, continuous variation in delay due to tape playback or sample rate variations, which has been addressed by using frame-by-frame cross-correlation to find the delay for each frame [Herre 2001]. Unlike speech quality measurement, the calibrated presentation level is provided as an input option, so an automatic level alignment process is not required.

The main processes in the filterbank auditory transform are shown in Figure 2.11. The processing is applied independently to each channel of the reference and degraded signals. The complex perceptual filterbank uses 40 bands, equally spaced on a bark scale at about 0.7 bark separation. The filters are designed for high frequency selectivity, using sinusoids modulated by a half-cycle squared cosine. Real and imaginary parts are calculated for each band using two filters with a 90° phase difference. The complex filter outputs are sub-sampled and passed through frequency-dependent gains to reproduce the linear filtering in the outer and middle ear. Frequency masking is modelled by spreading in the complex amplitude domain, with a level-dependent upper decay rate of –4 to –24dB/bark, and a constant lower rate of –31dB/bark. Performing this in the complex domain makes it possible to simulate the temporal response of deeply-overlapping perceptual filters, while allowing the filter shape to change with level.

The absolute square of the spread filter outputs is taken and smoothed in the time-domain with an 8ms half-cycle squared cosine, and then further sub-sampled to 4ms resolution. This first spreading models backward masking and the temporal resolution of the peripheral auditory system. Constant frequency-dependent noise is then added to mimic internal noise processes in the cochlea, and first-order low-pass filtering is used for a second time-domain spreading to model forward masking.

Figure 2.11: PEAQ filterbank auditory transform



Four different transforms are produced from the filterbank transform. (Two further transforms, the error signal and masked threshold, are only produced by the DFT model.) The amount of amplitude modulation in each band is computed from a Stevens sone loudness transform of the excitation prior to the second time-domain spreading. Specific loudness is computed using Zwicker's sone loudness formula, without a recruitment effect. The excitation after the second time-domain spreading is returned unprocessed in units of power. In addition, excitation is also returned after an adaptive two-stage equalisation process has been performed. Firstly, the levels are equalised. A copy of the excitation is low-pass filtered in the time-domain and used to estimate a correction factor (the inverse of the gain); if this is greater than one, the reference excitation is divided by that amount, otherwise the degraded is multiplied by it. Thus whichever signal is louder in the given frame is attenuated to the level of the other. After level correction, the frequency response is equalised in a broadly similar way, but with both low-pass filtering in the time-domain before phaseless transfer function estimation, and rectangular moving-average smoothing of the estimates in the bark domain. Again, whichever signal is louder in the given band is attenuated. This equalisation process is described further in section 4.4.2.1.

A total of fifteen distortion parameters are computed, known as model output variables (MOVs), of which eleven are used in the Basic version and five in the Advanced version. The MOVs fall into the following groups.

- averages of the difference in modulation between the two signals
- Zwicker sone noise loudness i.e. average positive error, weighted by an estimate of the masked threshold, and loudness of missing components i.e. average negative error
- amount of linear distortion or bandlimiting of the degraded signal

- mean noise-to-masked threshold ratio (NMR), and the mean rate in which NMR exceeds a threshold at any frequency in each frame
- probability of detection of differences, including binaural detection
- average peak harmonic error computed from the un-smoothed error in the DFT model.

These MOVs are mapped onto a prediction of subjective difference grade (SDG) [ITU-R BS.1116], known as objective difference grade (ODG). In both models a sigmoid neural network, with one hidden layer, is used. The Basic version, with three hidden nodes, has 40 free coefficients in the neural network, while the Advanced version, with five hidden nodes, has 36 free coefficients. The neural networks have no monotonic constraint, and the coefficient weights are fairly evenly spread between the two signs. The models were trained and tested on only a few hundred file pairs. Over-training was a major concern in the selection of BS.1387, so 32 file pairs from one database were held back from the developers and a new subjective test was conducted after training. These data were used to make the final selection of the output mappings.

More than four years after its standardisation, PEAQ is only used in a small number of laboratories, and no further independent studies of its accuracy have been found. BS.1116 listening tests are harder and more expensive to conduct than P.800 telephony tests, so there is much less validation data available. Any judgement on the performance of PEAQ must therefore be based on the selection results reported in [ITU-R BS.1387]. These suggest that the Basic model may be over-trained, in particular from the results for the 32 unknown file pairs from database DB3. Neither model fully met the performance requirement that was set out at the start of the competition: the Advanced model gives an ODG outside the desired tolerance in 34 of the 84 (i.e. 40%) of test file pairs – and 52 of these were known at the time of model training [ITU-R BS.1387]. Thus while PEAQ contains many interesting ideas it has not been shown to provide a highly reliably objective measure of subjective quality.

2.6 Performance evaluation

This section describes the methods that have become accepted in the ITU-T for the evaluation of the accuracy of perceptual models as predictors of subjective MOS, using subjective listening tests [Rix 1998d, Klaus 2000, Rix 2000a]. Unlike the evaluation of PEAQ described in the previous section, these methods are directed towards the analysis of large amounts of data, and address a number of characteristics specific to the P.800 ACR LQ tests that are not shared by BS.1116 audio tests.

2.6.1 Performance metrics

The following measures of prediction accuracy are most commonly used for the evaluation of perceptual models. (A full summary of the available statistical methods is outside the scope of this thesis.) In this thesis $y(i)$ is used to denote the subjective MOS for condition i in the given subjective test. $y_0(i)$ denotes the corresponding objective speech quality (OSQ), which may be on an arbitrary scale. $\hat{y}(i) = g(y_0(i),\mathbf{b})$ denotes the objective estimate of MOS, computed from $y_0(i)$ using some mapping function $g(\cdot)$ with coefficients $\mathbf{b}$, as discussed in sections 2.6.4 and 5.3 below.

Values averaged by condition are most commonly used for telephony. This is because a condition contains other sources of variation, in particular the talker and sentence material, which are not normally of interest [Rix 1998d]. However the same methods of performance assessment can be applied for each file [Rix 1998e].

2.6.1.1 Residual error

The residual error of the prediction is defined by equation (2-1). This is normally calculated after the mapping process i.e. from $\hat{y}(i)$, not $y_0(i)$.

$$\varepsilon(i) = \hat{y}(i) - y(i) \tag{2-1}$$

2.6.1.2 RMS error

A simple measure of performance is the root mean squared (RMS) residual error, shown in equation (2-2). This measures the average deviation in the same dimensions as the residual error. Minimum mean squared error (minimum MSE) is a common optimisation criterion and is used in Chapter 5.

$$RMSE = \sqrt{\frac{1}{N_i}\sum_{i=1}^{N_i}(\hat{y}(i) - y(i))^2} \tag{2-2}$$

2.6.1.3 Pearson's correlation coefficient

Pearson's correlation coefficient, or more simply, correlation coefficient, is a dimensionless measure of the strength of the relationship between two quantities [Duckworth 1968]. A value of zero indicates no relationship; a magnitude of 1 means that they are identical to within an offset and a scale. If the correlation is positive, the relationship is increasing; if negative, the relationship is decreasing or inverse. The correlation coefficient is as follows, where $\bar{y}$ is the mean of the set $\{y(i)\}$ and $\bar{\hat{y}}$ is the mean of the set $\{\hat{y}(i)\}$:

$$\varrho = \frac{\sum_{i=1}^{N_i} (\hat{y}(i) - \bar{\hat{y}})(y(i) - \bar{y})}{\sqrt{\sum_{i=1}^{N_i} (\hat{y}(i) - \bar{\hat{y}})^2 \sum_{i=1}^{N_i} (y(i) - \bar{y})^2}} \quad (2\text{-}3)$$

For typical functions used to compute $\hat{y}(i) = g(y_0(i), \mathbf{b})$, such as a linear or polynomial fit, if the coefficients **b** have been calculated to minimise RMSE according to equation (2-2), it can be shown that the mean error is zero, and hence

$$RMSE = \sigma_\varepsilon = \sigma_y \sqrt{(1 - \varrho^2)} \quad (2\text{-}4)$$

where σ_y is the sample standard deviation of $\{y(i)\}$ and σ_ε is the standard deviation of the zero-mean set $\{\varepsilon(i)\}$. (Equation (2-4) does not apply if $\{\varepsilon(i)\}$ is not zero mean.) In this case correlation coefficient can also be thought of as the square root of the proportion of the variance of $y(i)$ that is predicted by $\hat{y}(i)$ [Quackenbush 1988]. Thus ϱ^2=0.95 means that 95% of the variance of $y(i)$ is accounted for by the prediction; the remainder is the error variance.

2.6.1.4 Error distribution

RMSE provides a single estimate of the amount of prediction error, but gives no information about its distribution – in particular, the size and frequency of outliers. If the error sign is not of interest, a convenient summary of the error distribution is the cumulative distribution of absolute residual errors, either tabulated as in Table 2.2, or presented graphically. This shows that for this dataset, 96.6% of errors were less than 1.0. Alternatively, this can be plotted as a histogram of the raw or absolute residual error distribution.

Table 2.2: Example error distribution

Error magnitude e_t	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
Proportion less than e_t (%)	58.8	81.3	91.6	96.6	98.7	99.4	99.8	99.9

2.6.2 Other performance measures

While the measures listed in section 2.6.1 are in common use in the ITU-T for comparing perceptual models, many other measures can be considered.

2.6.2.1 Raw RMSE and correlation coefficient

As is shown below, it is often preferable to perform a mapping separately for each subjective test. However this approach neglects any systematic bias between OSQ $y_0(i)$ and MOS $y(i)$. An alternative method is to substitute $y_0(i)$ for $y(i)$ in equations (2-1), (2-2) and (2-3).

Equation (2-4) will not generally hold in this case. If the mapping from $y_0(i)$ to $\hat{y}(i)$ is linear, the correlation coefficient ϱ is unchanged.

2.6.2.2 Spearman's rank correlation coefficient

Pearson's correlation coefficient is highly dependent on the distribution of the data, in particular the extreme values of $\hat{y}(i)$ and $y(i)$. Non-linearity in the relationship will also tend to reduce the Pearson correlation. An alternative measure, Spearman's rank correlation coefficient, is equivalent to applying equation (2-3) with the rank order of each item in the sets substituted for $\hat{y}(i)$ and $y(i)$ [Duckworth 1968]. Ties (where several elements have the same value) must all be assigned the average of their ranks. This is equivalent to

$$r_S = 1 - 6\frac{\sum_{i=1}^{N_i} \Delta(i)^2}{N_i({N_i}^2 - 1)} \tag{2-5}$$

where $\Delta(i)$ is the difference of rank of element i in the sorted sets $\{\hat{y}(i)\}$ and $\{y(i)\}$. Like Pearson's correlation coefficient, this gives a value in the range [–1, 1], where a magnitude close to 1 indicates a good predictor. A weakness of Spearman's rank correlation coefficient is that it gives equal weight to rank order whether data points in $\{y(i)\}$ are widely spaced or are indistinguishable within experimental error. Since this is common in subjective test data, Pearson's correlation coefficient has been preferred.

2.6.3 Computational complexity

This thesis focuses on perceptual models for practical applications in speech quality measurement. The computational needs of these algorithms are therefore of significant interest. For the algorithms considered in this thesis, peak memory usage is proportional to the signal duration, and is no more than a few megabytes; this poses little problem for most of the applications envisaged and hence memory requirements are not considered further. Processing time is, however, very important, particularly for end-user applications in real-time network management, and for perceptual model development where it is necessary to process and perform training on a large database of material in a reasonable amount of time – over 25,000 reference and degraded file pairs were used for this thesis, as described in Appendix D.

To describe the computational complexity of algorithms in Chapter 3 and Chapter 5, Θ notation will be used. An algorithm that is $\Theta(N^2)$ requires, for large N, processing time proportional to N^2. Lower order terms, and multiplicative and additive constants, are neglected in the expression of Θ [Cormen 1990].

2.6.4 Comparison with subjective test data

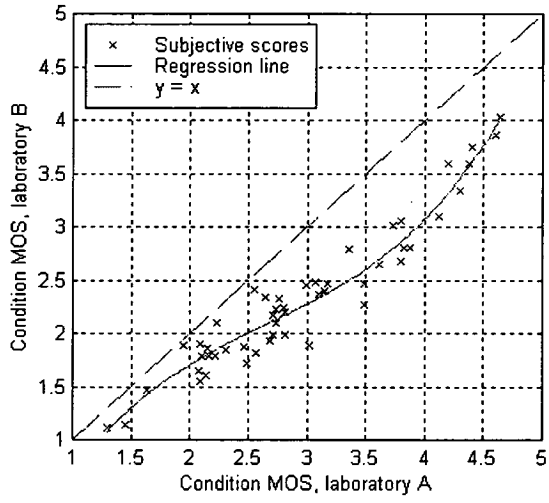
In general perceptual models may be calibrated on an arbitrary scale. This means that to compare objective quality score with MOS, in particular to calculate RMSE, some form of mapping or normalisation is required. As MOS varies from test to test, this mapping must be performed separately for each subjective test. This section introduces this variation and the methods used to address it. These issues are discussed further in Chapter 5.

2.6.4.1 Variation between subjective tests

A one-to-one comparison between subjective MOS from different subjective tests is difficult with tests conducted according to the ACR LQ method [ITU-T P.800, ITU-T P.830]. This is because subjective votes are affected by factors such as the following.

Cultural variation – in different languages and cultures, the meanings of "excellent .. bad" differ. This can have an effect of up to 1.0 MOS when comparing results for the same conditions from different laboratories. This is shown in Figure 2.12, which shows the scatter plot of conditions, and the 3rd-order polynomial regression line, between results of a test conducted in two different laboratories, each in their own native language and with native speakers as subjects. The network conditions used in the test were identical for each laboratory.

Figure 2.12: Cultural variation in MOS



Individual variation – users' personal experience also influences how they vote. As subjective tests tend to use a relatively small number of subjects (typically 24–32), systematic individual variations can leave a residual variation. Assuming that the standard deviation of individual variations is 0.5, with 24 subjects the 95% confidence interval for the mean quality score is on the order of 0.2 MOS.

Balance of conditions – the ACR method means that subjects adapt to some extent to the range of conditions in a test. If a large proportion of conditions are poor to bad, the best conditions are likely to be rated as excellent, as they are clearly distinct. Conversely, if there are fewer bad conditions, it is possible that the subjects may only rate the best conditions as good, as they are harder to distinguish. This can account for variations of up to 1.0 MOS between tests conducted at the same laboratory, which could not fully be explained by individual variations.

2.6.4.2 Normalisation

Previous work attempted to address these variations by agreeing an algorithmically-defined distortion, the modulated-noise reference unit (MNRU), and including this as a reference condition in every subjective test [ITU-T P.810, ITU-T P.861]. The MNRU adds noise modulated by the signal envelope at a specified SNR *Q*. A logistic function was then used to map subjective quality scores from MNRU to MOS for each test. By using the inverse mapping, quality for the conditions under test could be expressed in terms of equivalent *Q*. A similar process can be applied to map objective quality score onto equivalent *Q*.

The author found that although this method makes the MNRU conditions consistent across subjective tests, it does not reliably normalise other conditions. The MNRU is sensitive to the spectrum of the input speech signal, which varies according to the test design and the recording apparatus used in each subjective laboratory. In addition, the logistic function is not invertible outside its asymptotes, so conditions with MOS scores outside this range cannot be mapped onto equivalent *Q*. This subject is discussed further in section 5.4.7.

2.6.4.3 Comparison between objective and subjective quality

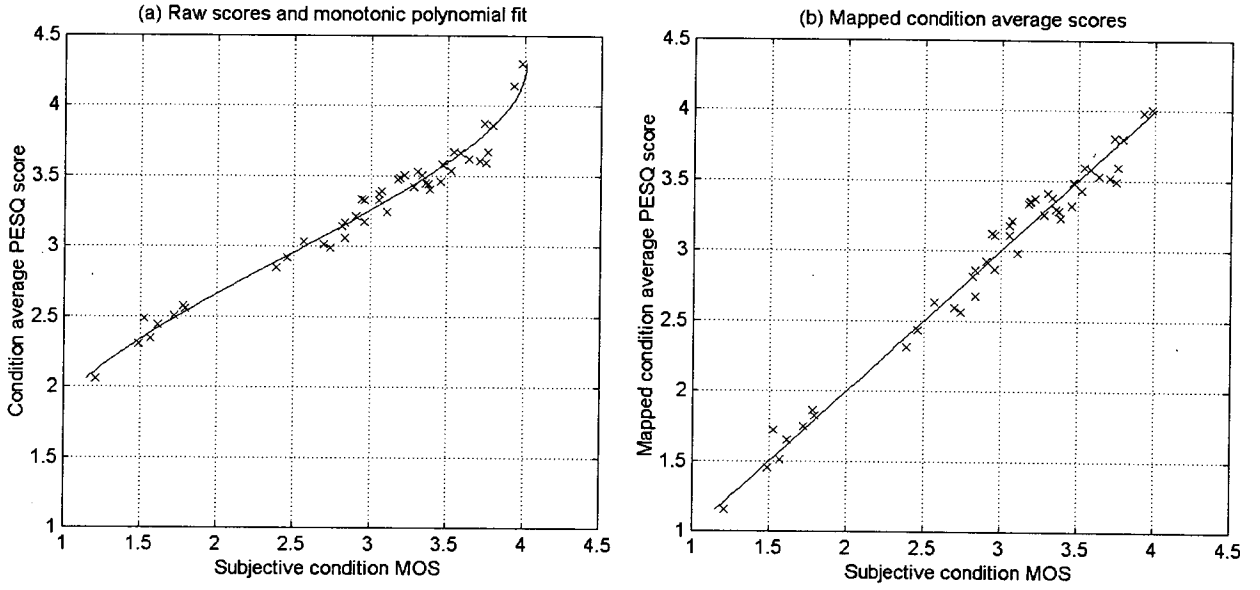
Although it is clear from the example shown above that subjective tests do not always give identical scores, it is reasonable to expect that rank order should be preserved within experimental error, and that there should be a monotonic relationship between MOS for equivalent conditions in two different tests. A monotonic mapping function can therefore be used to transform the results of one test onto the same scale as another. The same process must be used to compare objective speech quality (OSQ) against MOS: perceptual models are calibrated against some essentially arbitrary scale, but MOS varies systematically from test to test as shown by Figure 2.12.

Before this study, several different mappings were in use, making it difficult to compare results from different studies. The author developed a method using monotonic 3rd-order polynomial regression to compare OSQ and MOS, and this was adopted by the ITU for performance assessment of objective models. The algorithm used is detailed in section 5.3. This is applied, for each subjective test, to map the objective score onto the subjective score. It is then possible

to calculate correlation coefficient and residual errors. Usually the process is performed per condition, reducing material dependence, but it can also be applied per file.

This process is illustrated by the following example, a subjective test on the performance of fixed and mobile networks with errors, noise and noise suppression. Figure 2.13(a) shows a scatter plot between subjective MOS $y(i)$, on the x-axis, and PESQ score $y_0(i)$, along with the monotonic 3rd-order polynomial fit with minimum mean squared error. The PESQ score is mapped by this polynomial to give a prediction of subjective quality $\hat{y}(i)$, shown in Figure 2.13(b).

Figure 2.13: Mapping between objective and subjective MOS



2.7 Summary

This thesis aims to develop a perceptual model for assessment of the speech quality of telecommunications networks using intrusive measurements. This has many uses, from testing individual component devices or systems, to optimising or monitoring the quality of networks in service.

This scope of this study is much broader than that of earlier perceptual models, and is representative of the wide range of technologies in use in today's networks. At the points of connection to the network, either analogue or digital or (for the transmit side) acoustic connections may be used; noise may also be introduced at the send side. Many different signal processing and coding systems, such as VAD/DTX, speech codecs, and corresponding error or loss processes, may be encountered, including delay variation. A large database of subjective

test material has been assembled to evaluate the accuracy of perceptual models for these conditions. This range of applications means that conventional, linear signal processing techniques are difficult to use, posing a number of problems that this thesis will address.

Subjective testing provides a way to measure the performance of systems such as telephone networks, from the point of view of the customer. The ACR listening quality method is the most common, and gives a quality score termed mean opinion score (MOS) for each condition under test. However, subjective tests are slow and expensive to conduct, making them impractical for assessment of live networks. This provides the motivation for developing perceptual models that can be used to predict speech quality automatically.

Perceptual signal processing forms the basis of these models. The human auditory system has been extensively studied, and many large-scale effects such as the threshold of hearing, loudness perception and masking can be modelled algorithmically. Auditory transforms for estimation of the masked threshold are used in perceptual audio and speech coders.

The most common structure of perceptual models for quality assessment was first proposed by Karjalainen and is based on a comparison of auditory loudness transforms of the reference and degraded signals. The models PSQM, PAMS, PESQ and PEAQ that were introduced in this chapter all use this approach. The limitations of PSQM for conditions that included delay variation, linear filtering, or complex perceptual effects such as background noise, led to the developments by the author that are discussed in the following chapters.

The performance of a perceptual model is measured by comparing it against subjective test data. The main statistics that will be used in this thesis are correlation coefficient and RMS error. ACR subjective tests show large variations between tests, and monotonic mapping functions provide a means to compare objective perceptual models with subjective MOS. This will be used to evaluate the methods described in this thesis.

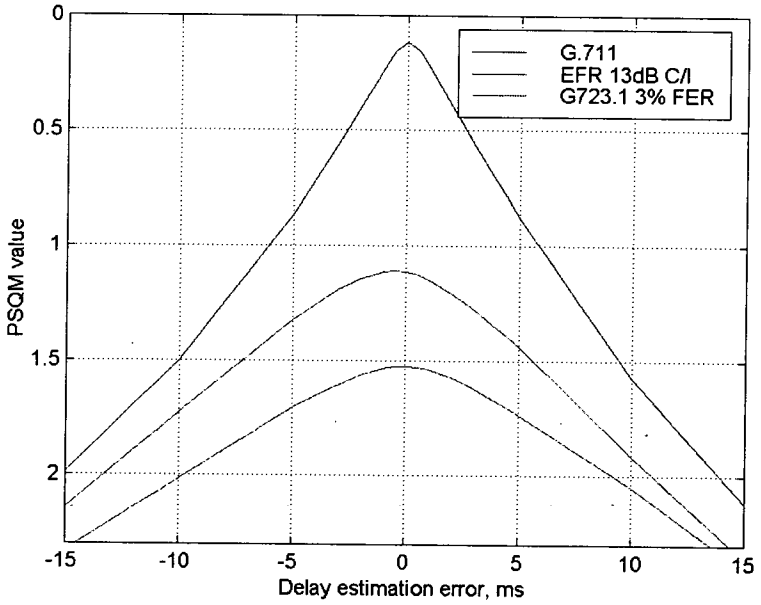
CHAPTER 3

Time delay identification for quality assessment

3.1 Overview

The perceptual models introduced in the previous chapter compute distance metrics from a frame-by-frame comparison of the transforms of the reference and degraded signals. As most audio signals are highly non-stationary, this comparison is meaningful only if the two signals are lined up. Even a small time offset can result in large false errors being observed. This is illustrated by Figure 3.1, where the PSQM distortion measure is shown against delay estimation error for an 8s measurement of a male talker with three simulated network conditions, all of which have known delay. For the G.711 condition, an increase of about 1.0 PSQM value – roughly a drop of 1.0 MOS – occurs if the delay estimate is 7ms in error.

Figure 3.1: Effect of incorrect delay estimation on PSQM value



Before the research described in this thesis began, it was acknowledged that time alignment was important in perceptual quality assessment. P.861 states that for "signals that show group delay distortion, the delay that leads to the minimum PSQM value is the correct one" [ITU-T P.861, section 9.1.1]. To solve this adequately would require slow iterative application of the entire perceptual algorithm, requiring at least that the transform of either the reference or degraded signals be evaluated at every possible delay, increasing the complexity of the algorithm by a factor of 64 at 8kHz sampling rate. In practice, P.861 suggests using single-step cross-correlation to estimate delay. This is a faster but less robust solution, as is shown in section 3.3.

A more detailed study was published by Tallak et al [Tallak 1993], who compared six methods for the estimation of constant delay: single-step cross-correlation; cross-correlation of (full rate) envelopes; cross-correlation of zero-crossing-rate estimates; cepstral distance; mean spectrogram difference SNR; and coherence-based signal-to-distortion ratio, also evaluated from the spectrogram. This found that signal cross-correlation was unsatisfactory and that the two spectrogram-based comparisons gave the greatest accuracy but, as with optimising delay for minimum PSQM value, these are very computationally intensive.

To address the need for a fast but reliable method for delay identification, the author developed a two-stage algorithm combining envelope-based crude delay identification followed by fine alignment using a method based on a histogram of short-term cross-correlation delay estimates. An alternative derivation using a maximum a posteriori (MAP) approach was also developed by the author, and leads to a similar method. Both solutions are presented in section 3.4, and both are found to be more robust to non-linear processing than cross-correlation, but are of similar computational complexity.

Like their predecessors, the histogram and MAP algorithms assume constant time-delay in the system under test. During the course of this research, systems which break this assumption became increasingly common. In particular new voice-over-packet networks, such as voice over IP (VoIP), have been found to exhibit frequent variations in delay. To address this problem, the author extended the delay identification algorithm to recursively estimate the delay of sections of the signals, using a confidence measure to identify changepoints. Section 3.5 describes this process and compares it to dynamic time-warping, which has been used by Herre [Herre 2001] and Beerends for the same purpose.

This extended algorithm was implemented in the PAMS perceptual model in 1998–1999, over a year before any competing models offered this functionality. The algorithm was key to the success of PAMS in the ITU-T P.862 competition, and was integrated with a further extension by Hekstra into PESQ [ITU-T P.862]. PESQ is used to illustrate the influence of time-delay estimation on perceptual models in the examples presented in sections 3.4.4 and 3.6.

3.2 Background and assumptions

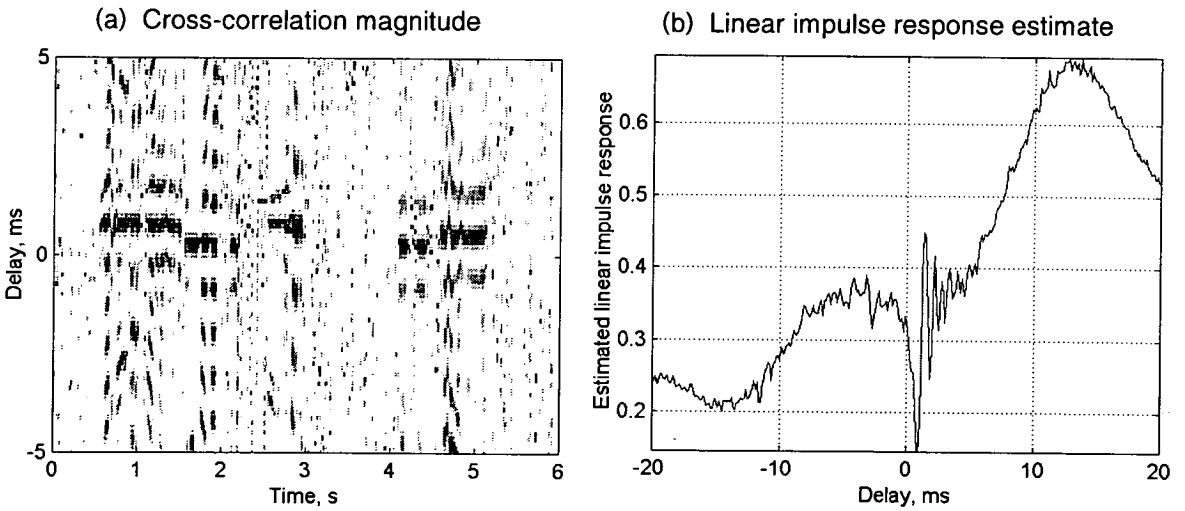
Before methods for time-delay identification are introduced, this section defines the key assumptions that are required by some of these algorithms.

3.2.1 Linear, time-invariant system

The theory of system identification for LTI systems is well-established and forms the basis of several of the classical techniques introduced below. However this thesis focuses on telecommunications systems that potentially combine low bit-rate codecs, coding errors, and other distorting processes such as sampling rate clock or packet jitter. These systems are both non-linear and time-variant. This is the main cause of problems with methods such as linear transfer function estimation or cross-correlation that essentially compute the average impulse response across the whole measurement.

This is illustrated by Figure 3.2, which is taken from a low bit-rate mobile condition with coding and radio errors, in a subjective test conducted by the author. Figure 3.2(a) shows an image of the magnitude of cross-correlation of the two signals, with pre-filtering, on 64ms 75% overlapping frames using a Hann window. Each frame is normalised to the same maximum amplitude; dark values indicate strong cross-correlation. The location of the maximum cross-correlation gives an estimate of the delay, suggesting that the delay varies between about 0–1ms due to the coder and channel errors. However, the linear impulse response computed using equation (3-1), plotted in Figure 3.2(b), shows a strong maximum at 13ms delay. If this were chosen, the delay estimate would be about 12ms in error.

Figure 3.2: Effect of time-variation on linear methods



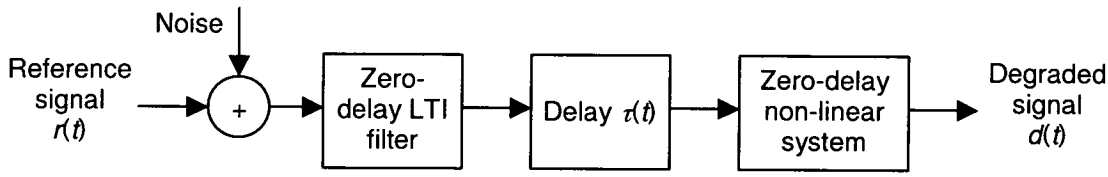
3.2.2 Decomposition of the system

Excluding non-linear components such as codecs, telecommunications networks often include filtering between the measurement points. Most commonly, this occurs because measurement includes the 2-wire local loop or the electro-acoustic path of one or both telephone handsets. These components are normally LTI. For the algorithms presented in sections 3.4 and 3.5, it is therefore assumed that the system can be decomposed into the following stages, shown in Figure 3.3.

- Additive noise
- Acausal LTI filter of zero mean delay
- Simple delay, which may vary during the measurement
- Zero-delay, time-varying, instantaneous non-linear system.

This decomposition is required only for time-delay identification, and is not intended to be a full description of the system. The time-delay identification problem can be specifically stated as the identification of only the simple delay component.

Figure 3.3: System decomposition for time-delay identification

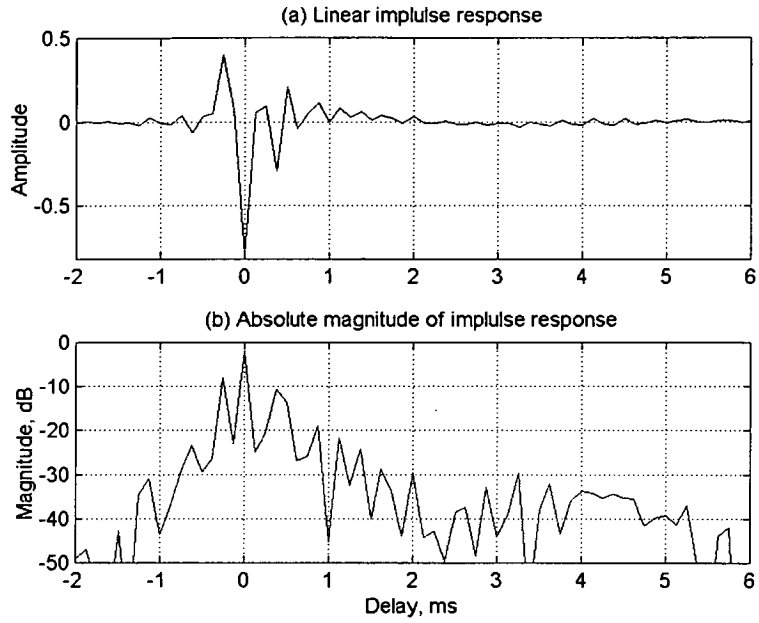


3.2.3 Dispersion

Frame-based delay identification methods such as the histogram-based method that will be introduced in section 3.4 make an implicit assumption that the dispersion of any filtering in the network is small compared to the perceptual model's frame duration. There are various ways in which dispersion can be quantified – for example, the difference between minimum and maximum group delay in the passband, or the shortest interval within which 90% of the impulse response energy falls.

The electrical and electro-acoustic components of telecommunications networks are usually simple, resulting in an impulse response duration much shorter than 1ms, whereas the frames used for time alignment are normally 16ms or longer. This is illustrated by Figure 3.4, which shows the impulse response of a PSTN handset, including both acoustic and electrical paths, shown in amplitude and in dB units. The impulse response was estimated using the method described in section 4.3.2.2, with a 512-point FFT at 8kHz sampling rate, using 75% overlapping Hann windows. The gain falls below –10dB within ±0.5ms, and below –20dB within ±1.0ms.

Figure 3.4: Impulse response of handset



3.2.4 Delay variation

The first part of this chapter describes methods that assume constant time-delay. A consequence of this assumption is that certain processes which result in delay variation will be classified as of constant delay, and may result in significant errors in estimation of the mean delay. Some examples include the following:

- Sample rate jitter or variation in record/playback rates, for example with analogue tape. These can cause continuous variations in delay [Herre 2001]
- Packet loss concealment by resampling. Some algorithms slow down the playback of packets as jitter buffers near empty, causing the local timebase, and potentially the pitch, to distort [Liang 2001]
- Re-synthesising parametric codecs. Some very low bit-rate codecs using MELP, VXC or other vocoder methods can result in very significant changes to the pitch or the number of pitch cycles in each speech event. This results in large timebase variations that cannot be modelled by a constant delay.

In the second part of this chapter, section 3.5, time alignment for variable delay systems is explored. The assumption made here is that the system is piecewise constant: delay changes are instantaneous and discrete. This assumption is justified for systems such as VoIP where the delay change is a result of a buffer over- or under-flow, and delay changes are relatively

infrequent. The assumption is found to be necessary because the scope of conditions of interest make it impossible to robustly identify delay variation on short timescales.

3.2.5 Perceptual effect of delay and delay variation

Based on the effect on perceived speech quality in a listening context, the different classes of delay variation can be summarised as follows.

Delay changes in silence are normally inaudible. In certain cases the system may introduce some type of discontinuity, for example a short period of digital silence corresponding to an increase in delay, but even this will usually have little perceptible effect. Large delay changes may have some impact on conversational quality, but this is outside the scope of this thesis.

Step delay changes during speech are normally audible. Even a small change in delay can cause an annoying discontinuity in phase or amplitude. However, some error concealment algorithms attempt to hide delay changes, for example, by inserting or deleting one complete pitch cycle during a voiced part of speech. Perceptual processing to evaluate this is discussed in section 3.5.7.

The effect of continuous delay changes during speech depends strongly on how the temporal structure of speech is affected. If the pitch is made constant, for example by very low bit-rate speech coders, the resultant "robotisation" of the speech is sometimes found by subjects to be annoying. However, as long as some natural variation in pitch remains, it can be difficult to detect any impairment. Based on informal listening tests, continuous warping of the time axis by less than about 1% per second was found by the author to be inaudible with speech signals.

3.3 Existing techniques

A number of methods for identifying delay are summarised in this section, drawing on techniques from system identification and sensor array processing. A number of the concepts for linear system identification that are used here are described in further detail in section 4.3.

The bulk of the literature in this area has focused on estimation of the time delay between a signal that is received, with noise, at pairs of independent sensors. Applications include range and bearing estimation and, for techniques that can identify continuously-varying delay, velocity estimation. One of the most widely cited procedures is Knapp and Carter's generalized cross-correlation method, which is discussed in section 3.3.3. This provides a maximum-likelihood (ML) delay estimate in the presence of uncorrelated noise at the sensors [Knapp 1976], and unifies a number of techniques in a framework of filtering the cross-spectrum estimate. More recent surveys and frameworks may be found in [Carter 1981, Meyr 1984, Johnson 1993, Brandstein 1995, Carlemalm 1997, Stuller 1997].

3.3.1 Time-delay identification from the impulse response

Under the assumption that the impulse response is known or can be identified, and the dispersion is low, a good estimate of the time-delay may often be taken by locating the absolute peak in the impulse response $h(t)$. Absolute magnitude is used because sign is sometimes not preserved in telephone networks, as was the case in Figure 3.4.

This method was described by Roth, using cross-spectrum-based transfer function estimation to compute a nonparametric estimate of the impulse response according to equation (3-1) [Roth 1971, Söderström 1989].

$$\hat{h}(t) = \mathrm{F}^{-1}\left[\frac{P_{rd}(\Omega)}{P_{rr}(\Omega)}\right] \tag{3-1}$$

Here $P_{rd}(\Omega)$ and $P_{rr}(\Omega)$ are the cross-spectrum and auto-spectrum estimates, and F^{-1} denotes the inverse Fourier transform. This approach exhibits good rejection of noise in $d(t)$ and can be efficiently evaluated using Welch's modified periodogram method [Welch 1967], although the original formulation [Roth 1971, Knapp 1976] used the entire signals. The complexity of Welch's method is $\Theta(N_r \log N_f)$, where N_r is the number of samples in $r(t)$ and N_f is the frame length used, while cross-correlation of the entire signals using the FFT is $\Theta(N_r \log N_r)$, which is typically a factor of seven larger for the applications considered in this thesis. The periodogram-based method is discussed in more detail in section 4.3.2 in the context of transfer function estimation. As discussed in section 3.3.4, the periodogram requires that the absolute delay is much less than N_f, and this affects the complexity of this method.

Parametric methods may also be applied to estimate the impulse response directly, for example using least squares [Chan 1980], which is discussed further in section 4.3.1. Alternatively, adaptive filters may be used [Reed 1981, Feintuch 1981, Carlemalm 1997]; the most common is the NLMS algorithm because of its simplicity, stability and (in appropriate conditions) fast convergence [Haykin 1996]. [Carlemalm 1997] post-processes the filter weights by a Kalman filter to produce smoothed estimates, and uses this to perform significance tests to identify which tap represents the delay. For sub-sample estimation accuracy, [Chan 1980] suggested interpolation using the sinc function; other authors have performed quadratic fitting about the maximum in $\hat{h}(t)$ [Boucher 1981]. Subspace methods such as MUSIC have been proposed for multi-sensor, multi-source partitioning and bearing estimation [Johnson 1993].

The complexity of these parametric methods, typically $\Theta(N_r N_h)$, depends on the search range N_h over which $\hat{h}(t)$ is computed. Like the cross-spectrum method, this must include all possible candidate delays. As the complexity rises with N_h rather than log N_f, these methods are generally more computationally intensive than the cross-spectrum method.

3.3.2 Frequency-domain delay estimation

A simple system with gain a and constant delay of τ samples has z-transform $az^{-\tau}$, corresponding to the frequency response shown in (3-2).

$$H(e^{j\Omega}) = ae^{-j\Omega\tau} \tag{3-2}$$

From the theory of linear systems, the frequency response at a given frequency can be fully described by the gain a and phase shift, in the case of this pure delay $-\Omega\tau$. By considering the derivative of the phase of $H(e^{j\Omega})$, the delay at radian frequency Ω can therefore be calculated using (3-3). This relationship holds for general LTI transfer functions $H(e^{j\Omega})$ [Oppenheim 1989].

$$\tau(\Omega) = -\frac{\mathrm{d}}{\mathrm{d}\Omega}\angle H(e^{j\Omega}) \tag{3-3}$$

An example of the group delay estimated using (3-3) for a simulated linear system is shown in Figure 3.6.

An early ad hoc approach that uses the phase of the cross-power spectrum only to estimate a pure delay was the phase transform [Knapp 1976] (3-4)

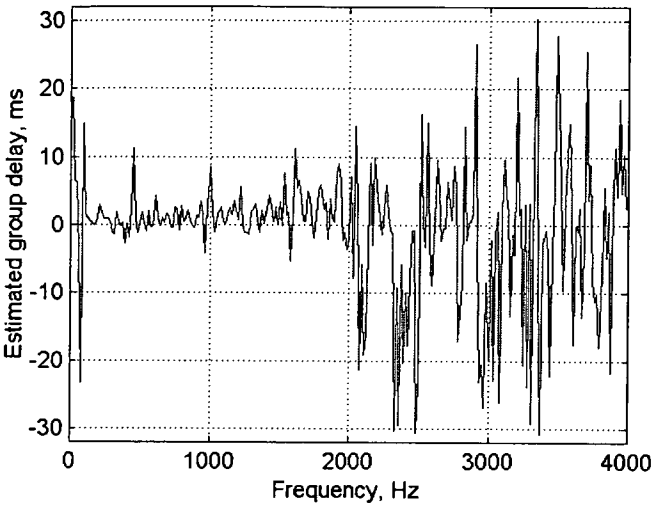
$$\tilde{H}(e^{j\Omega}) = \frac{P_{rd}(\Omega)}{|P_{rd}(\Omega)|} \tag{3-4}$$

More recently, a family of methods for delay identification using the cross-spectrum phase have been investigated by Brandstein, who has examined time-delay estimation for speech sources received at multiple microphones, for source location estimation [Brandstein 1995, Brandstein 1997]. The general process here is to perform a linear fit on the unwrapped cross-spectrum phase $\angle P_{rd}(\Omega)$.

A recurring problem for the application of these techniques within the scope of this thesis is the following: the cross-power spectrum, and hence the phase, is often not a very stable function and is particularly sensitive to coding errors, noise, and periods of severe distortion.

This is illustrated by the example shown in Figure 3.5. This is based on the same measurement as Figure 3.2, a low bit-rate mobile condition with noise at the talker. From the signal cross-correlation in Figure 3.2 it appears that the delay is in the range 0–1ms. Figure 3.5 shows the group delay calculated using (3-1) and (3-3), with unwrapping of phase shifts greater than π. The transfer function estimation in both Figure 3.2 and Figure 3.5 used the 512-point FFT at 8kHz sampling rate, with 75% overlapping Hann windows.

Figure 3.5: Group delay jitter



3.3.3 Signal detection by cross-correlation

Carter and Knapp modified Roth's delay estimation method by generalising the cross-correlation function between signals [Knapp 1976]. They considered the case where the system between the two signals is non-dispersive, with arbitrary spectra for the signal and noise.

The starting-point for this approach is the theorem that the cross-correlation between two signals, and the cross-spectrum between them, form a Fourier transform pair (3-5) [Söderström 1989]. In the discrete-time notation of this thesis, the generalised cross-correlation (3-7) may be evaluated from the cross-spectrum (3-6), with a positive frequency-dependent weighting $\psi(\Omega)$. The value of t that maximises (3-7) is taken as the estimate of delay $\hat{\tau}$.

$$\mathrm{E}[r(t+\tau)d(t)] = \int_{-\pi}^{\pi} \mathrm{E}[R^*(e^{j\Omega})D(e^{j\Omega})]e^{j\Omega\tau}\,\mathrm{d}\Omega \tag{3-5}$$

$$P_{rd}(\Omega) = \mathrm{E}[R^*(e^{j\Omega})D(e^{j\Omega})] \tag{3-6}$$

$$\hat{h}(t) = \int_{-\pi}^{\pi} \psi(\Omega)P_{rd}(\Omega)e^{j\Omega t}\,\mathrm{d}\Omega \tag{3-7}$$

In the above, E[] is the expectation operator. These results hold for quasi-stationary signals [Söderström 1989] and can be applied to speech using a windowed method.

It can be seen that if $\psi(\Omega)=1/P_{rr}(\Omega)$, (3-7) reduces to the Roth (transfer function) estimate shown in equation (3-1). Carter and Knapp showed that the maximum likelihood delay estimator $\hat{\tau}$ in the presence of noise on both measurements $r(t)$ and $d(t)$ is produced by

$$\psi(\Omega) = \frac{C_{rd}(\Omega)}{|P_{rd}(\Omega)|(1 - C_{rd}(\Omega))} \tag{3-8}$$

where $C_{rd}(\Omega)$ is the magnitude squared coherence between $r(t)$ and $d(t)$ given in equation (4-13).

Depending on whether windowed methods or the entire signals are used in the evaluation of $P_{rd}(\Omega)$, the complexity of these methods are either $\Theta(N_r \log N_f)$ or $\Theta(N_r \log N_r)$, as discussed in section 3.3.1.

3.3.3.1 Simplifications of cross-correlation

Many authors have used cross-correlation to estimate the time-delay with no frequency-dependent processing, for example [Tallak 1993, ITU-T P.861]. This is equivalent to using $\psi(\Omega)=1$ in Carter and Knapp's method, or $P_{rr}(\Omega)=1$ in Roth's method.

If $r(t)$ has a flat spectrum, the factor $P_{rr}(\Omega)$ in equation (3-1) is constant and can therefore be ignored for the purpose of time alignment. This is the case if, for example, $r(t)$ is independent, identically distributed (IID) or white noise, or an appropriately designed signal such as a chirp or maximum-length sequence, or equivalently by (3-5), its autocorrelation function approximates an impulse. This reduces the problem to computation of the inverse Fourier transform of the cross-power spectrum $P_{rd}(\Omega)$ of the signals, as shown in (3-9).

$$\hat{h}(t) \approx \mathrm{F}^{-1}[P_{rd}(\Omega)] \tag{3-9}$$

Alternatively, this can be posed as a time-domain cross-correlation (3-10) using the result of (3-5):

$$\hat{h}(\tau) = \mathrm{E}[r(t+\tau)d(t)] \approx r(-t) * d(t) \tag{3-10}$$

where * denotes convolution. In both of these methods, the location of the absolute maximum of the estimate of $h(t)$ gives an estimate of the delay. (3-9) and (3-10) can be evaluated efficiently with complexity $\Theta(N_f \log N_f)$ and $\Theta(N_r \log N_r)$ respectively using the FFT. Computation of (3-10) requires rather more memory than the windowed periodogram methods that may be used to estimate $P_{rd}(\Omega)$; conversely, the windowed method requires the signals to be approximately aligned, a problem avoided by cross-correlating the whole signals.

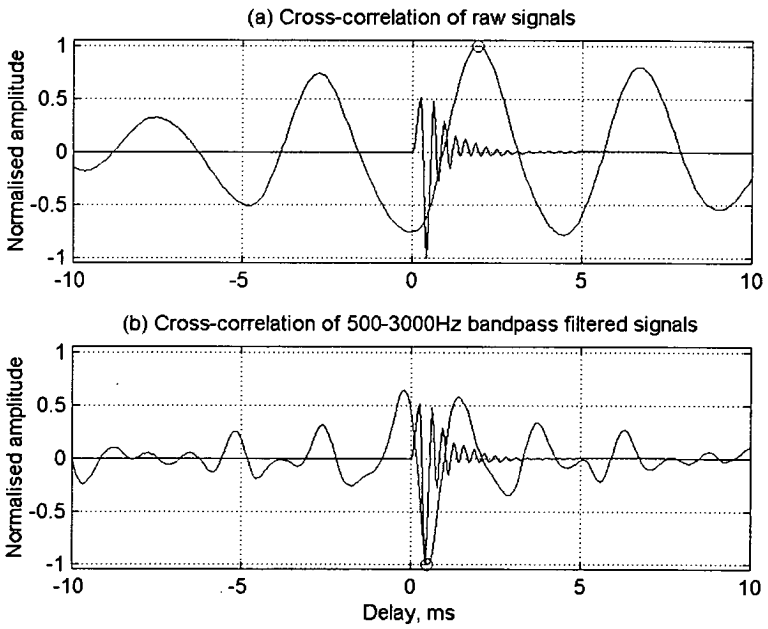
3.3.3.2 Bias due to non-linear phase

Models such as PSQM that have used simple cross-correlation have often ignored the fact that $P_{rr}(\Omega)$ is not constant. Real signals such as speech and music are highly coloured, often with most energy at low frequencies. Because telecommunications or audio systems are designed to transmit these signals, it is highly desirable to use test signals that are representative of them

– for the purpose of this thesis, either natural speech or artificial speech-like test stimuli [Hollier 1995]. However, the spectral colouration of these signals can result in estimation bias for systems with non-linear phase. The result is that the delay estimate tends to be biased towards the frequencies in $r(t)$ that contain the greatest energy.

This is illustrated by the following example, which is based on an IIR filter implemented by the author to model the IRS send response [ITU-T P.48] of a standard telephone handset (see section 4.2.2 and Figure 4.22). An 8s sentence pair spoken by a female talker, at 16kHz sampling rate, was used. The filter impulse response, and the signal cross-correlation from equation (3-10), are plotted in Figure 3.6(a). The "true" delay, measured by the location of the absolute peak of the impulse response, is at 0.44ms; however, the absolute maximum of the cross-correlation is at 1.94ms, marked by the circle in Figure 3.6(a). This is 1.5ms in error. Furthermore, it can be seen in Figure 3.6(a) that there are several subsidiary maxima, due to the autocorrelation of $r(t)$, of magnitude within 75% of that of the cross-correlation peak; in the presence of noise or other distortions, one of these could be picked, potentially leading to an even greater error in the delay estimate.

Figure 3.6: Group delay bias in cross-correlation



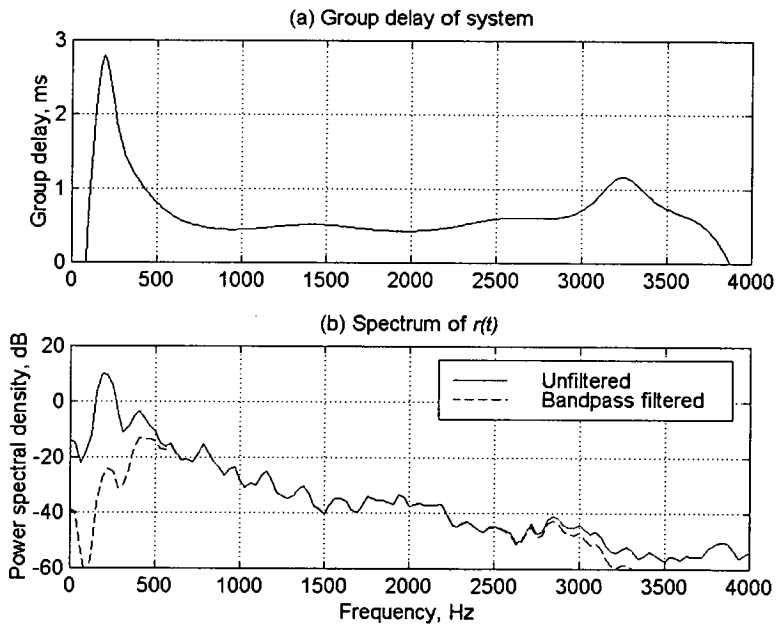
A general approach to controlling this bias is to pre-filter the signals $r(t)$ and $d(t)$. Knapp and Carter envisaged applying $\psi(\Omega)$ by filtering either or both of the signals being compared; the product of the two frequency responses must be $\psi(\Omega)$ for the cross-correlation to be maximum likelihood [Knapp 1976]. It should be noted, however, that the formula of (3-8) makes the problem even worse in this case, as the coherence is close to unity at all frequencies, but the

factor $1/|P_{rd}(\Omega)|$ gives too much weight to frequencies below 200Hz and above 4kHz, where the group delay is very different from the passband. This is because (3-8) was derived for non-dispersive systems.

The control of bias by pre-filtering is shown by Figure 3.6(b), where both $r(t)$ and $d(t)$ have been filtered by a 4th-order Butterworth bandpass filter with –3dB frequencies 500 and 3,000Hz [Oppenheim 1989]. The filtered cross-correlation gives a delay estimate of 0.5ms, only one sample different from the impulse response maximum. The subsidiary maxima are also reduced both in amplitude and distance from the true delay. The reason for this difference is illustrated by Figure 3.7, which plots the group delay of the IRS filter and the spectrum of the original, and bandpass filtered, reference signal. The unfiltered speech signal has most energy between 150–250Hz, where the group delay is above 2ms. However, the bandpass filtered signal is strongly attenuated below 400Hz, leading to a delay estimate that is much closer to the mean group delay in the passband.

The method of pre-filtering to control group delay bias is used in the delay estimation algorithm described in the next section.

Figure 3.7: Group delay and spectral bias



3.3.4 Crude delay estimation

Both transfer function estimation and the windowed cross-correlation method are only able to identify delay within a fraction of the frame length used for the windowed FFT (typically within $\pm N_f/4$ samples for an N_f-point FFT). Although in telecommunications applications the duration of

the impulse response is normally within a few milliseconds, the bulk delay may be a second or more, while the frame length is typically on the order of 50ms. It is therefore necessary to estimate the delay to within some fraction of N_f samples; the appropriate offset is then used to eliminate the estimated delay from the calculation of $h(t)$. Two alternative methods are commonly used to achieve this:

- signal matching – locating a distinctive signal component such as a chirp, for example by a matched filter method
- correlation of signal envelopes.

An envelope-based crude delay estimation algorithm is described in the next section.

The accuracy of this crude delay estimation becomes less critical as N_f is increased. However, this increases computational complexity and reduces the noise rejection and stability of the transfer function estimate. The author found that a good balance between these factors was achieved for speech quality measurement by a frame length N_f of 64ms.

3.4 Histogram-based method for delay estimation

This section describes the method that the author developed for time alignment in PAMS, and which was also incorporated into PESQ [ITU-T P.862]. The core of the method is the computation of a weighted histogram of delay estimates for each time alignment frame, smoothed with a kernel function. Pre-processing is performed to make this method more robust and to make it an approximately maximum likelihood procedure. The histogram approach is compared to a Bayesian maximum a posteriori (MAP) derivation. Results are presented in section 3.4.4 to illustrate the choice of constants and to compare the performance of this method with standard cross-correlation based methods.

3.4.1 Pre-processing

As introduced in section 3.3.3, cross-correlation based methods give delay estimates that are weighted by the spectral content of $r(t)$. Window-based methods are also limited to delay values that are within a fraction of the frame length. Two pre-processing stages are performed to control these effects.

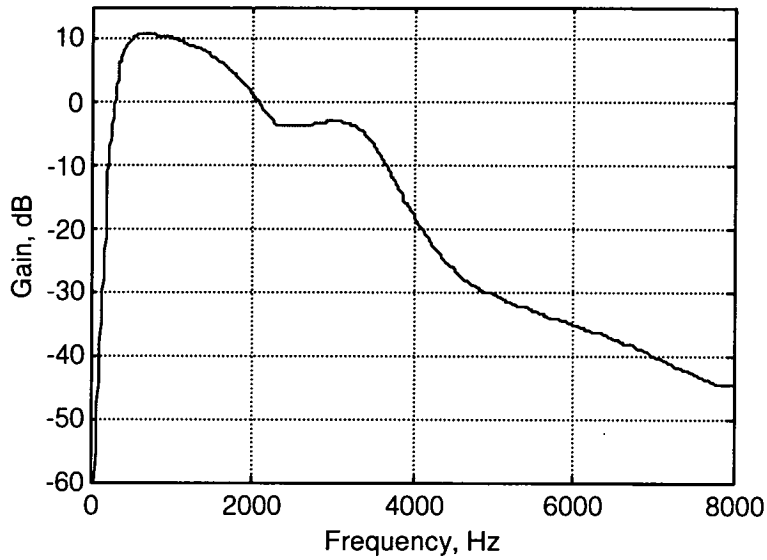
3.4.1.1 Filtering

In PAMS, both $r(t)$ and $d(t)$ are processed by the same input filter, which models the frequency response of a narrowband telephone handset and the leakage of the ear canal. Time alignment is therefore performed on the signals as they would be heard by the user. The frequency response of this filter is shown in Figure 3.8; essentially, the filter falls off strongly below 300Hz,

and has a peak in response in the region 1–2kHz. This substantially reduces the effect of the low frequency content of $r(t)$ on the delay estimate, and acts to make the cross-correlation delay estimate closer to the maximum-likelihood estimate of Knapp and Carter (section 3.3.3).

In PESQ, this processing is performed after the PESQ input filter. This filter has a band-pass characteristic, attenuating strongly below 250Hz and above 3500Hz, and is flat from 600–3250Hz. It was found that this additional pre-filtering improved the accuracy of the time alignment by further reducing the low frequency bias effect.

Figure 3.8: Frequency response of time alignment input filter



3.4.1.2 Crude delay estimation

Envelope-based cross-correlation has been proposed to derive an estimate of the bulk delay. Voran used a power envelope-based method for MNB [Voran 1999a], while Tallak found that the low-pass filtered amplitude envelope gave good delay estimates [Tallak 1993]. For the method considered here, the envelope-based delay estimate is used to roughly align the signals before subsequent processing, to keep within the capture range of the window-based method.

A consequence of this is that if the crude delay estimate is outside this range, it will not be corrected later. The author discovered that two key properties of the test signals and system have a significant impact on the accuracy of the crude delay estimate, and used these to improve the envelope-based method.

The envelope of noise in either signal (or both) may correlate well with the main speech signal, and the noise is unknown for the scope of this thesis. If there is heavy distortion to the speech, such as temporal clipping, combined with impulsive or speech-like noise, the cross-

correlation between the noise and the speech may exceed that between the speech in the two signals, leading to an incorrect delay estimate. Typically this will result in large errors in the delay estimate.

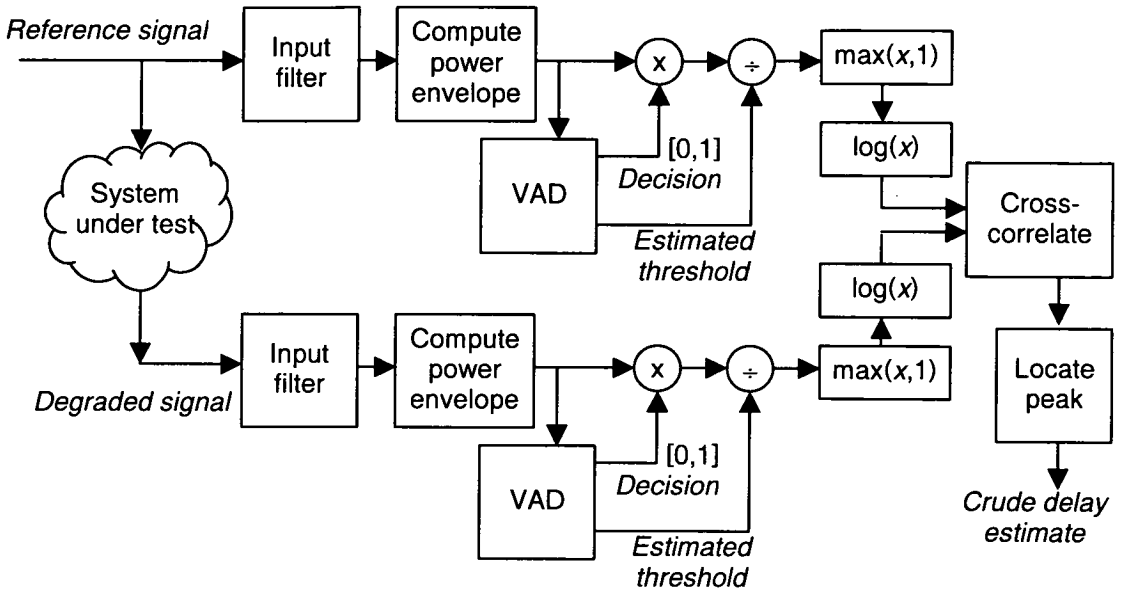
Loud distortions in $d(t)$ may also correlate well, in the envelope domain, with the main speech signal. If a distortion is much louder than the speech in $d(t)$, a false maximum in cross-correlation may be found, typically matching the loudest part of $r(t)$ with the loudest distortion. These distortions may occur during speech, so the delay estimation error may be of any value and cannot be completely eliminated by setting some maximum delay threshold.

The author addressed these issues for PAMS by making two modifications to the envelope computation. The first was to use voice activity detection (VAD) to eliminate the effect of noise in silent intervals; the second was to perform the alignment using the log of the envelope divided by the threshold, reducing the weight given to loud distortions.

This processing is shown in Figure 3.9. After input filtering, the crude delay algorithm begins by computing the power envelope, on rectangular non-overlapping frames of 4ms duration. This envelope is used as input to the VAD, which is described below. The VAD decision is used to set the envelope to zero in silent periods, so they have no weight in the cross-correlation. The endpoint accuracy of the VAD was found not to be critical, although it is essential that if the VAD is ambiguous, no zeroing should be performed. Most subjective tests of telephone networks with noisy speech use either a good SNR (above 10dB), or use noise signals such as vehicle noise that are predominantly low-frequency and are highly attenuated by the filtering described in the previous section, resulting in similarly good SNR, so VAD ambiguity has been found to be a minor issue.

The envelope is then divided by the VAD threshold estimate, and the log is taken. For periods where the envelope has been zeroed or is below threshold, the input is set to one, so that the log is zero. The processed log envelopes of the two signals are cross-correlated, and the index of the maximum is used as the crude delay estimate. Compared to using the amplitude or power envelopes, the log envelope was found to reduce the bias due to loud distortions.

Figure 3.9: Crude delay estimation algorithm

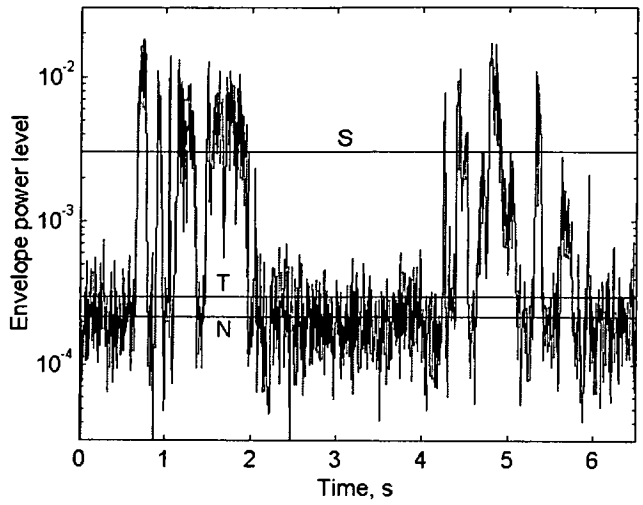


The VAD that the author used is similar to that described by [Van Compernolle 1990], with one main modification. While Van Compernolle calculated the Gaussian minimum error threshold, the author found that the Gaussian distribution was not a good model for the envelope level of speech, and found better performance with a simpler threshold estimate based solely on the noise mean plus two standard deviations, computed in the power domain using an iterative approach.

The VAD envelope and thresholds are illustrated in Figure 3.10, which shows the degraded file from a mobile network connection with noise at the talker and radio errors. The VAD threshold is initially set at the mean power of the signal; envelope frames above this threshold are classified as speech, and below as noise. An iteration is performed to update the estimates of the speech and noise levels, and re-set the threshold, until convergence is reached. Additional processing is performed to label short events, that are above threshold for 12ms or less, as non-speech, and to join together speech utterances that are separated by less than 200ms of silence.

The full algorithm is provided in functions apply_VAD() and crude_align() in pesqdsp.c in [ITU-T P.862]. Optimum values for constants such as the frame length, threshold update and section durations were found during the development of PAMS in 1997–98.

Figure 3.10: VAD decision threshold



Key:

S Estimated active speech level

N Estimated noise level

T Estimated threshold, set to noise level plus two standard deviations in units of power

The operation of the crude delay estimation process is shown in Figure 3.11 and Figure 3.12. Figure 3.11 plots the log envelope of the reference and degraded signals. This illustrates that the envelope is set to zero during silent intervals, and shows the elevation of threshold in Figure 3.11(b) due to the noise in the degraded signal. Figure 3.12(a) shows the result of cross-correlating the envelopes, and Figure 3.12(b) presents the centre of the cross-correlation function, showing a clear peak at 0.280s that gives the estimate of the delay.

Figure 3.11: Crude delay log envelopes

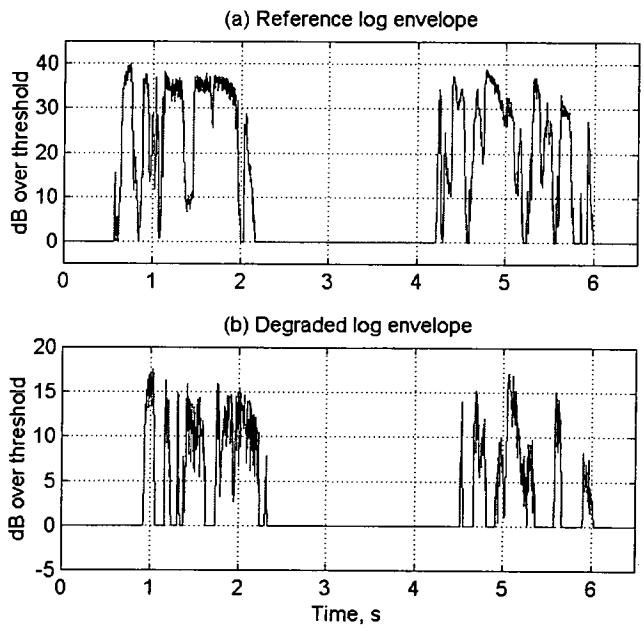
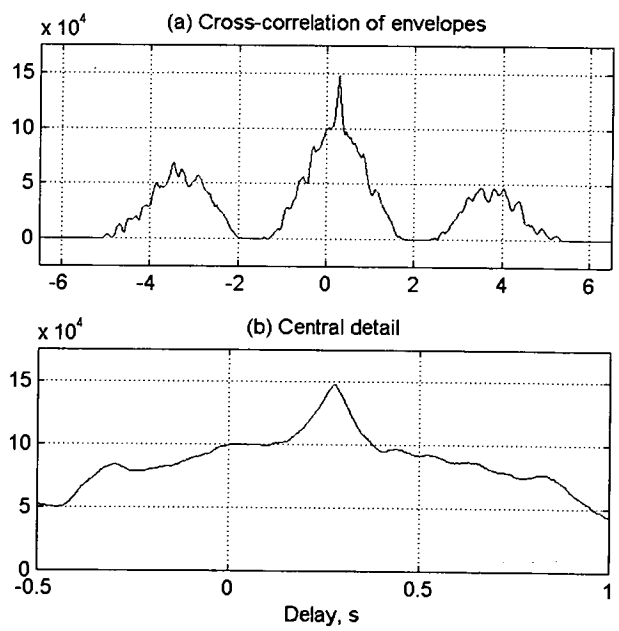


Figure 3.12: Envelope cross-correlation



3.4.2 Weighted delay histogram

This section presents a new method that the author developed for estimating a probability density function (PDF) for the time delay. The use of a histogram for this purpose was first suggested during discussions with colleagues at BT Labs, but it extends a concept that had been considered by Meyr and Spies [Meyr 1984], who performed short-term cross-correlation and then applied a modified Kalman filter to compute a smoothed estimate of time-varying delay. The approach developed by the author focuses instead on the estimation of a constant delay by constructing a histogram of frame delay estimates, smoothed using a kernel method [Bernardo 2000], and weighted by a function of the magnitude of the frame cross-correlation. A Bayesian derivation of a similar algorithm that has been produced by the author is presented in the next section.

Following [Meyr 1984], the starting point is to compute a maximum-likelihood cross-correlation delay estimate on short windowed frames of the signal, to ensure that time-varying effects such as clock jitter will be negligible. These delay estimates are assembled into a histogram, which will be subject to noise due to two separate processes.

- Severe distortion is likely to produce large, randomly distributed delay estimates, causing scatter that will have little effect on the main peak of the histogram apart from reducing its height.
- Non-linear phase systems, combined with time-varying spectral content of the signals (see section 3.3.3.2) will tend to produce localised spreading of the delay estimates about the "true" delay. Peak picking in this case is ambiguous, motivating the use of kernel smoothing [Bernardo 2000].

A disadvantage of the use of a histogram is that all frames have equal weight. Quieter frames, such as during silent periods, are most likely to be corrupted by noise or suppressed in DTX, and should therefore be given reduced weight. Conversely, it is undesirable to give too much weight to a small number of very loud sections, as these may be subject to distortion. The author therefore extended the method to weight the values used to construct the histogram.

A simple power law, applied to the peak value of cross-correlation in each frame, was found to address both of these aspects in a satisfactory way. Many different powers were tested during the development of PAMS, and an exponent of 0.125 was found to produce greatest overall accuracy. In the absence of distortions, the cross-correlation peak is in units of energy, so this exponent may be compared to perceptual loudness scales that apply a similar power law to signal intensity, with exponents from 0.23 [Zwicker 1990] to 0.3 [Stevens 1972]. The use of a compressive function is compared to the evaluation of log likelihood in the next section, and the optimum weights and kernel size are evaluated in section 3.4.4.

3.4.2.1 Algorithm

The full algorithm for computing the smoothed, weighted delay histogram is as follows. It is assumed that the signals have been pre-filtered and roughly aligned as described in section 3.4.1.

The signals $r(t)$ and $d(t)$ are divided into 75% overlapping frames of length N_f samples, and windowed using the Hann raised cosine window. The DFT is used to evaluate the cross-correlation function (3-11)

$$\chi(\tau,k)=\sum_{t=0}^{N_f} w(t+\tau)r(t+\tau,k)\cdot w(t)d(t,k) \tag{3-11}$$

where $w(t)$ is the window function, $r(t,k)$ and $d(t,k)$ are the reference and degraded signals at time t, $0\le t<N_f$, in frame k. In practice the FFT method used, on N_f points, actually performs circular convolution. The window function minimises the circular effect for $|\tau|<<N_f$, provided that there is little correlation between the signals when wrapped around. It was not found to give any improvement to the accuracy of the perceptual model to use zero padding to evaluate (3-11) exactly, but this would more than double the computational complexity of the algorithm.

The delay estimate for the frame, $\hat{\tau}(k)$, is taken from the index of the absolute maximum of $\chi(\tau,k)$ (3-12).

$$\hat{\tau}(k)=\arg\max_{\tau}\left|\chi(\tau,k)\right| \tag{3-12}$$

The weighted histogram is computed by assembling the frame delay estimates with a weight calculated from $\chi(\tau,k)$ using a power law with an exponent α=0.125 (3-13), where $\delta(t)$ is the discrete Dirac delta function [Oppenheim 1989].

$$p_h(\tau)=\sum_k \delta(\tau-\hat{\tau}(k))\left|\chi(\hat{\tau}(k),k)\right|^{\alpha} \tag{3-13}$$

Kernel smoothing is applied by convolution of the histogram with a kernel function, as shown in (3-14). In this case the kernel function $\kappa(\tau)$ is symmetric and centred about τ=0. In the development of PAMS it was found by the author that the triangular kernel shown in (3-15), where the half-width β=1ms, led to good overall accuracy.

$$p_s(\tau)=\kappa(\tau)*p_h(\tau)=\sum_t \kappa(t)p_h(\tau+t) \tag{3-14}$$

$$\kappa(\tau)=\begin{cases}1-\dfrac{|\tau|}{\beta}, & |\tau|<\beta\\ 0, \text{otherwise}\end{cases} \tag{3-15}$$

For computation, it is most efficient to evaluate (3-13) over all frames, and implement the smoothing as a post-process using (3-14). However, for comparison with the maximum-likelihood method derived in the next section, it is convenient to formulate the smoothed weighted delay histogram directly by carrying the convolution into the sum and using the properties of the delta function (3-16):

$$p_s(\tau) = \sum_k \varkappa(\tau - \hat{\tau}(k)) \left| \chi(\bar{\tau}(k), k) \right|^{\alpha} \tag{3-16}$$

The location of the peak of the smoothed histogram $p_s(\tau)$ is chosen as the delay estimate $\hat{\tau}$. This is added to the crude delay estimate to give the estimate of the total delay of the system.

3.4.2.2 Confidence measure

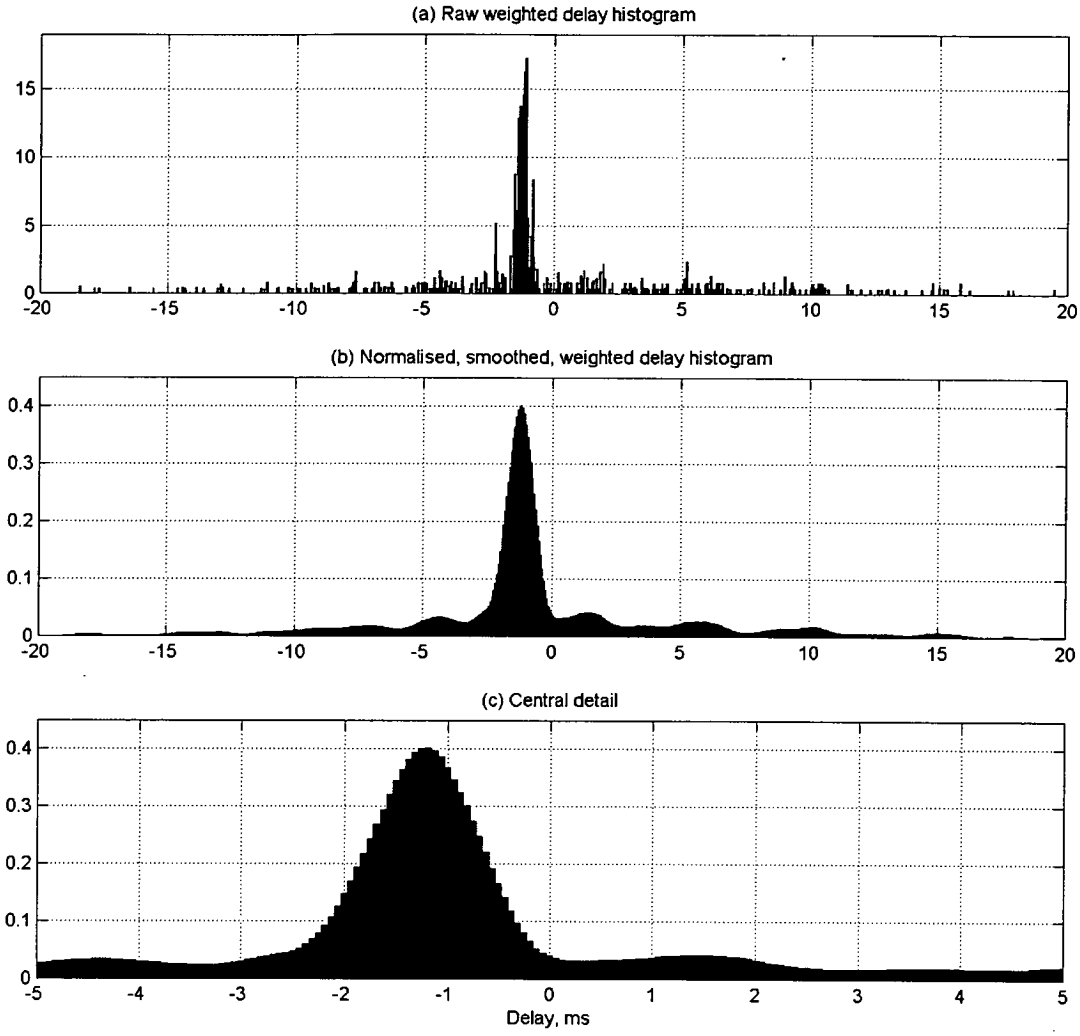
By normalising the histogram to sum to 1 prior to kernel smoothing, the magnitude of the peak may be used as a confidence measure related to the consistency of the delay estimate. This means that if all frame delays fell within the same histogram bin, the confidence would be 1; if they were evenly distributed across the frame (with constant weight), the confidence would be $1/N_f$, where N_f is the frame length (512 samples at 8kHz sampling rate). As discussed below, this confidence measure was found to be particularly useful in variable delay alignment, and was much more robust than using a waveform-based measure such as the cross-correlation at the estimated delay.

C code that implements this algorithm, including kernel smoothing and confidence estimation, is provided in function time_align() in pesqdsp.c in [ITU-T P.862].

3.4.2.3 Example smoothed, weighted delay histogram

The result of applying the method outlined in this section to the example condition introduced in section 3.4.1.2 is shown in Figure 3.13. The frame length used here is 1024 samples (64ms) at 16kHz sampling rate. The fine delay estimate found here is –1.19ms, giving a total delay estimate for this test case as 278.81ms.

Figure 3.13: Smoothed weighted delay histogram



3.4.2.4 Complexity

The main processing operation in this method is the evaluation of the frame cross-correlation (3-11). Assuming that the frame size and overlap are constant, the complexity of the histogram method is therefore $\Theta(N_r \log N_f)$, where N_r is the number of samples in $r(t)$ and N_f is the cross-correlation frame size. This is within a constant multiplier of the cost of evaluating the cross-spectrum using Welch's method: in practice the latter requires only one inverse FFT, while (3-11) requires an inverse FFT every frame, in addition to the two forward FFTs that both methods require per frame. The complexity of the histogram method is therefore only 1.5 times greater than the most efficient implementations of the standard time-delay estimation algorithms described in section 3.3.

3.4.3 Bayesian approach to delay estimation

The approach described in the previous section was developed empirically. The author also produced a theoretical derivation of a maximum a posteriori (MAP) estimate of the delay, which is presented here and leads to a structure similar to the histogram-based algorithm. This approach is based on Bayes' theorem (3-17) [Bernardo 2000].

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{3-17}$$

A common interpretation (see for example [Rajan 1997]) is that $P(A)$ represents prior knowledge about the parameter A, and $P(B|A)$ represents the "evidence", or likelihood of the data B given the particular value of A. $P(B)$ serves as a normalising constant to ensure that the posterior probability distribution $P(A|B)$ sums to 1 over all possible values of A. Thus (3-17) makes it possible to combine prior knowledge with data to produce an updated estimate of the probability distribution of the parameter. The MAP method chooses the value of A that maximises (3-17).

3.4.3.1 Formulation for a frame of data

Like the approach shown in the previous section, following [Meyr 1984], this technique is applied frame-by-frame. It is assumed that the frame size is sufficiently short that time-varying distortions such as clock jitter or speech re-synthesis will either have minimal effect over each frame or, in the event of severe localised distortion, will tend to randomise the delay estimate.

Two cases are considered. In the presence of severe distortion, represented by D, we have no further information about the delay and all delays in the search range will be treated as equally likely. However, in the absence of distortion (represented by $\tilde{D}$), the likelihood of the data given a candidate delay can be evaluated; furthermore, the data can be used to provide information on the probability of distortion $P(D)$. Denoting a candidate delay by τ, and the data for the given frame by k, Bayes' theorem gives (3-18):

$$p(\tau \mid k, \tilde{D}) = \frac{p(k \mid \tau, \tilde{D})p(\tau \mid \tilde{D})}{p(k \mid \tilde{D})} \tag{3-18}$$

where

$$p(k \mid \tilde{D}) = \int_{\tau} p(k \mid \tau, \tilde{D})p(\tau \mid \tilde{D})\,\mathrm{d}\tau \tag{3-19}$$

which serves as a normalising constant. The equivalent formulation for non-distorted frames can be obtained by substituting D for $\tilde{D}$ in (3-18) and (3-19). The posterior probability density for delay for frame k is then derived by summing over $\tilde{D}$ and D:

$$p(\tau \mid k) = p(\tau \mid k, \tilde{D})p(\tilde{D} \mid k) + p(\tau \mid k, D)p(D \mid k) \tag{3-20}$$

The component probability distributions that make up equations (3-18) and (3-20) will now be considered.

With no other knowledge the prior probability distributions of delay τ, in the absence of distortion, $p(\tau \mid \tilde{D})$, or in the presence of distortion, $p(\tau \mid D)$, are both taken to be uniform:

$$p(\tau) = a_1 \tag{3-21}$$

Similarly, in the presence of distortion, there is no information about the likelihood of the frame at any one delay compared to another:

$$p(k \mid \tau, D) = a_2 \tag{3-22}$$

Combining (3-21) and (3-22) using (3-18) and (3-19), where $\tilde{D}$ is replaced by D, and substituting constant a for the reciprocal of the integral of τ, gives a non-informative posterior distribution on τ in the presence of distortion:

$$p(\tau \mid k, D) = a \tag{3-23}$$

In the absence of distortion, it is assumed that a maximum likelihood delay estimate $\hat{\tau}(k)$ can be found using (3-12). As discussed in section 3.4.1.1, pre-filtering must have been performed to ensure that this will, in appropriate conditions, be equivalent to the maximum likelihood delay estimate for that frame. Given this delay estimate, a two-sided exponential function with parameter b is used to model the probability density of the data given a delay τ, representing uncertainty due to distortion, dispersion, and bias caused by time-varying spectral content of the signal with non-linear phase (3-24).

$$p(k \mid \tau, \tilde{D}) = \frac{1}{2b} \exp\left(\frac{-|\tau - \hat{\tau}(k)|}{b} \right) \tag{3-24}$$

The two-sided exponential function was chosen as long tails were found by the author in the distributions of $\hat{\tau}(k)$ for many network measurements, as illustrated in the example presented in Figure 3.13(a); these would be poorly modelled by assuming a Gaussian distribution. This approach could be extended by using a power other than unity on the argument of the exponential, or other long-tailed probability distributions.

The posterior probability in the absence of distortion is obtained by substituting (3-24) and the uniform prior (3-21) into (3-18) and (3-19) to give (3-25).

$$p(\tau \mid k, \tilde{D}) = p(k \mid \tau, \tilde{D}) = \frac{1}{2b} \exp\left(\frac{-|\tau - \hat{\tau}(k)|}{b} \right) \tag{3-25}$$

The remaining task to allow (3-20) to be evaluated is to estimate the probability of distortion from the data.

In the absence of distortion ($\tilde{D}$), the likelihood of a given frame is taken as a constant times the absolute cross-correlation at $\bar{\tau}(k)$ (3-11), which is proportional to the covariance of the signals in that frame. This is used to model the probability of frame k given $\tilde{D}$ (3-26):

$$p(k \mid \tilde{D}) = c \cdot \left|\chi(\bar{\tau}(k), k)\right| \tag{3-26}$$

In the presence of distortion, we have no knowledge, so again a uniform prior will be assumed, to give $p(k \mid D) = a_3$. Prior knowledge about the likelihood of distortion, $p(D)$, can be incorporated using Bayes' theorem (3-17), to give

$$p(\tilde{D} \mid k) = \frac{c \cdot \left|\chi(\bar{\tau}(k), k)\right| \cdot p(\tilde{D})}{p(k)} \tag{3-27}$$

and

$$p(D \mid k) = 1 - p(\tilde{D} \mid k) = \frac{a_3(1 - p(\tilde{D}))}{p(k)} \tag{3-28}$$

where

$$p(k) = c \cdot \left|\chi(\bar{\tau}(k), k)\right| \cdot p(\tilde{D}) + a_3(1 - p(\tilde{D})) \tag{3-29}$$

Finally, (3-27), (3-28), (3-23) and (3-25) can be substituted into (3-20) to give the estimate for the posterior probability distribution of delay for the frame (3-30)

$$p(\tau \mid k) = \frac{c \cdot \left|\chi(\bar{\tau}(k), k)\right| \cdot p(\tilde{D}) p(t \mid k, \tilde{D}) + a \cdot a_3(1 - P(\tilde{D}))}{p(k)} \tag{3-30}$$

(3-30) can be simplified by absorbing the prior knowledge $p(\tilde{D})$, the normalisation $p(k)$, and the constant a_3 into constants a and c to give

$$p(\tau \mid k) = a\left[c \cdot \left|\chi(\bar{\tau}(k), k)\right| \cdot \frac{1}{2b} \exp\left(\frac{-\left|\tau - \bar{\tau}(k)\right|}{b}\right) + 1\right] \tag{3-31}$$

This represents the posterior probability of delay τ at frame k given the prior knowledge represented by the constants b and c. a provides normalisation only.

3.4.3.2 Maximum a posteriori delay estimate

The log-likelihood of delay $L(\tau|k_1, k_2,\ldots)=L(\tau|\mathbf{k})$ for all frames is derived from (3-31) by taking the log-product over k (3-32). The maximum-likelihood delay estimate $\hat{\tau}$ is given by the value of τ that maximises $L(\tau|\mathbf{k})$.

$$L(\tau \mid \mathbf{k}) = \log\left[\prod_k p(\tau \mid k)\right] = N_k \log a + \sum_k \log\left[\frac{c}{2b}\exp\left(\frac{-|\tau - \hat{\tau}(k)|}{b}\right)\cdot|\chi(\hat{\tau}(k),k)| + 1\right] \tag{3-32}$$

(3-32) can be used to evaluate the posterior distribution of delay τ, with normalisation (3-33).

$$p(\tau \mid \mathbf{k}) = \frac{\exp(L(\tau \mid \mathbf{k}))}{\sum_t \exp(L(t \mid \mathbf{k}))} \tag{3-33}$$

A simple estimate of the likelihood of the estimated delay given the data may be computed by integrating $p(\tau|\mathbf{k})$ over an appropriate interval about $\hat{\tau}$.

3.4.3.3 Interpretation and comparison with the smoothed weighted delay histogram

Two parameters in (3-32), b and c, may affect the maximum posterior likelihood. (As noted above, the third parameter, a, serves as a normalising constant.) The interpretation of b and c may be made clear by comparing (3-32) with the equivalent formulation of the smoothed weighted delay histogram (3-16). Separating out the individual components for frame k from (3-32) and (3-16) respectively gives:

$$L'(\tau \mid k) = \log\left[\frac{c}{2b}\exp\left(\frac{-|\tau - \hat{\tau}(k)|}{b}\right)\cdot|\chi(\hat{\tau}(k),k)| + 1\right] \tag{3-34}$$

$$p_s'(\tau \mid k) = \kappa(\tau - \hat{\tau}(k))|\chi(\hat{\tau}(k),k)|^{\alpha} \tag{3-35}$$

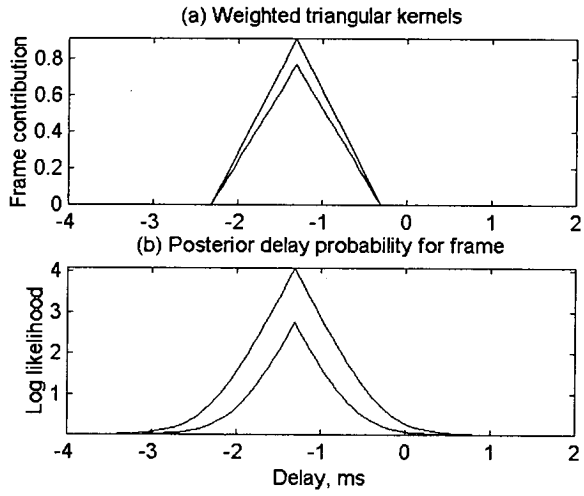
It is apparent from these equations that the histogram is comparable to a log-likelihood probability density function, despite the use of $p()$ notation in section 3.4.2.

Parameter b is analogous to the kernel slope parameter β (3-15). However, the effect of the additive constant 1 in (3-32), combined with $\frac{c}{2b}|\chi(\hat{\tau}(k),k)|$, is to set a variable width on the smoothing implemented by (3-32).

Examples of (3-35) and (3-34) are plotted in Figure 3.14 for two frames with the same $\hat{\tau}(k)$ but different values of $|\chi(\hat{\tau}(k),k)|$: 0.46 and 0.12. The parameters for the histogram method are α=0.125 and β=1ms, and for the MAP method are b=0.25ms and c=1000. While the kernel method uses constant width set by β, the MAP method has maximum slope determined by b but

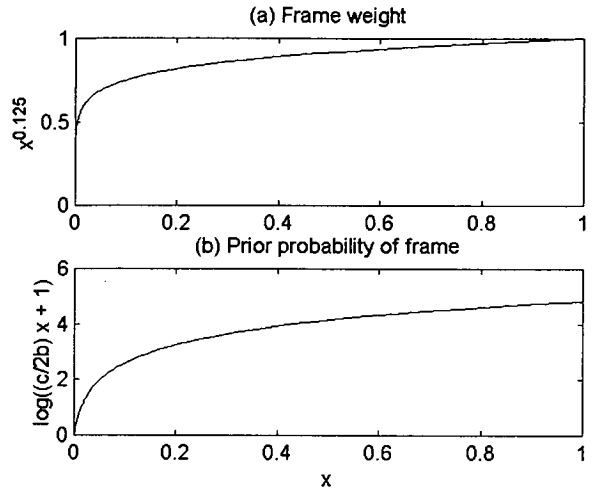
effective width determined by $\frac{c}{2b}\left|\chi(\tilde{\tau}(k),k)\right|$, and the base of the log probability density function for the MAP method curves smoothly to an asymptote of zero, while the kernel used in (3-15) has a simple triangular shape. This is a consequence of the choice of evidence probability density $p(k \mid \tau, \tilde{D})$ (3-24).

Figure 3.14: Comparison of histogram and MAP delay methods



The other main difference between these two approaches is the use of a power law to derive the frame weight in the histogram method (3-35), where a logarithmic relation is used in the MAP method (3-34). The relationship with $\left|\chi(\tilde{\tau}(k),k)\right|$, for values in [0, 1] and with $\tau = \tilde{\tau}(k)$, is shown in Figure 3.15. While both functions are compressive, the power law gives greater curvature close to zero. Thus the MAP method, with these values of b and c, gives slightly less relative weight to frames with low cross-power. This also accounts for the difference in the relative peak heights shown in Figure 3.14. The choice of c is dependent on the magnitude of $\left|\chi(\tilde{\tau}(k),k)\right|$, and does control the shape of the curve: for small values of $c\left|\chi(\tilde{\tau}(k),k)\right|$ the relationship becomes much more linear than in Figure 3.15(b), as $\lim_{x\to 0} \log(x+1) = x$.

Figure 3.15: Comparison of histogram and MAP weight functions



3.4.3.4 Complexity

The main processing operation in the MAP method is the evaluation of (3-11) to compute the frame cross-correlation. More operations are performed per frame to evaluate the log likelihood than the corresponding histogram weight. However, assuming that efficient implementations (such as look-up tables) are available for of the log() and exp() functions, (3-11) dominates and the algorithm complexity is $\Theta(N_r \log N_f)$, which is the same as for the histogram method.

3.4.3.5 Further development

Strictly, the posterior probability density on delay should be written as $p(\tau|\mathbf{k},\mathbf{b},M)$, rather than $p(\tau|\mathbf{k})$. This formulation allows τ to be generalised to a vector of estimates of delay and other quantities – the next section extends the histogram method to the identification of multiple piecewise constant delays and their changepoints. This formulation also specifies the parameters that affect the density, shown by vector **b**, which corresponds to b and c in the derivation above. Finally, the structure and order of the method is represented by model M, allowing different models to be compared in this Bayesian framework. Numerical methods could then be applied to marginalise the parameters, taking into account any prior knowledge, to compute τ and/or to identify the most likely of several models, without the need to assume particular parameter values. However, it should be noted that these methods are much more computationally intensive than the direct estimation procedure developed in this section.

3.4.4 Results

This section evaluates the different delay estimation algorithms by measuring their effect on the performance of PESQ, with appropriate modifications, over the subjective test database

described in Appendix D. The correlation coefficient, calculated per condition using monotonic 3rd-order mapping for each subjective test, as described in section 2.6, is used as a measure of the model's accuracy.

3.4.4.1 Optimisation of histogram method

The choice of the constants N_f, α and β for the weighted smoothed histogram method was investigated by altering these in PESQ. The default values, which were originally found during the development of PAMS for constant-delay applications, are N_f=64ms, α=0.125 and β=1ms.

The effect of the frame size N_f was determined by adjusting the constants Align_Nfft_8k and Align_Nfft_16k in pesqpar.h [ITU-T P.862]. All other delay estimation processes, including utterance splitting and bad frame realignment, were unchanged (these are discussed in the next section). Table 3.1 summarises the model performance for values of 32, 64 and 128ms. This shows that the highest overall performance is found with N_f=64ms. Although the worst-case performance might suggest that N_f=128ms would be slightly better, this value did lead to a drop of 0.4% in the correlation compared to N_f=64ms, for one live mobile network subjective test.

Table 3.1: Effect of delay estimation frame size on model performance

Frame size N_f, ms	32	64	128
Mean correlation, constant-delay tests	0.9547	0.9550	0.9549
Worst-case correlation, constant-delay tests	0.9012	0.9007	0.9008
Mean correlation, all tests	0.9417	0.9435	0.9435
Worst-case correlation, all tests	0.8115	0.8108	0.8120

The correlation weight power α and kernel size β were found to interact with the bad frame realignment process described in section 3.5.6, which effectively acts to correct earlier mistakes in delay estimation. Bad frame realignment was therefore turned off for the evaluation of these constants using PESQ. This makes a small reduction in overall model performance.

Table 3.2 presents results on the influence of α. From this data it appears that a choice of α=0.25 would be best in terms of overall and worst-case performance; examining the results for individual tests shows that in one case the correlation is improved by 0.5% compared to α=0.125. However, in other cases and overall, the default setting of α=0.125 derived for PAMS appears to be close to optimum. Values below 0.125 or greater than 0.5 appear to cause a substantial drop in overall performance, reducing correlation for some individual tests by as much as 5%, which suggests that the correlation weighting process makes a very significant contribution to the accuracy of the time-delay estimate.

Table 3.2: Effect of correlation weight power on model performance

Correlation weight power α	0.0625	0.125	0.25	0.5	1	2
Mean correlation, constant-delay tests	0.9545	0.9547	0.9547	0.9544	0.9544	0.9538
Worst-case correlation, constant-delay tests	0.8994	0.8994	0.8993	0.8994	0.8998	0.8982
Mean correlation, all tests	0.9416	0.9430	0.9434	0.9430	0.9418	0.9398
Worst-case correlation, all tests	0.8057	0.8055	0.8071	0.8053	0.8056	0.8056

The action of kernel smoothing was investigated by altering the kernel width β, with bad frame realignment disabled. Figure 3.16 shows the effect of β for the constant-delay subjective tests, showing what is in fact a very weak relationship with overall model performance (less than 0.02% difference in the mean, and 0.1% in the worst-case correlation).

Including variable-delay subjective tests, Figure 3.17 shows a slightly different relationship. Here the mean performance increases as the kernel size grows, due to two variable-delay subjective tests that give best performance with values of 12ms, rising 2.3% and 0.8% from the performance with the default value of 1ms. The worst-case performance is slightly reduced for β>2ms, by up to 0.2%, but this is small compared to the improvement gained for the variable-delay tests. This data does suggest that the choice of β=12ms, making the kernel much stronger than the default width of 1ms, would give better performance in variable-delay conditions.

Figure 3.16: Effect of kernel width on model performance, constant-delay tests

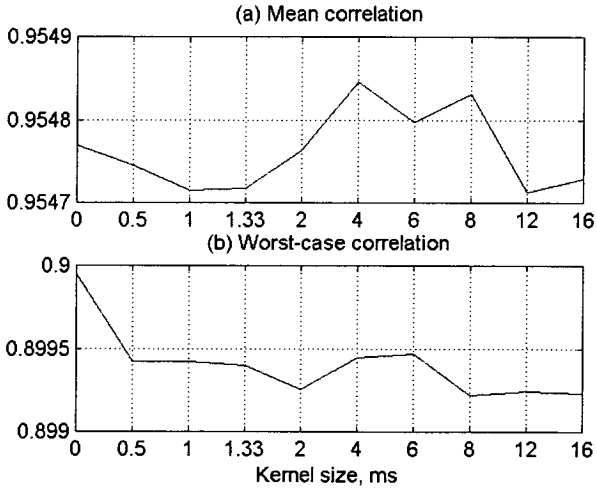
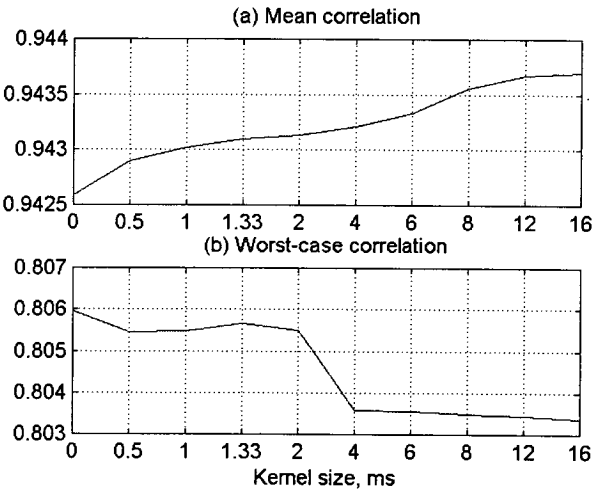


Figure 3.17: Effect of kernel width on model performance, all tests



3.4.4.2 Comparison of algorithms

For constant-delay subjective tests, the performance of the histogram method may be directly compared with cross-correlation. In PESQ, this was achieved by disabling the utterance

identification, splitting and bad frame realignment processes that are described in the next section, to perform the delay assessment over the entire signals using either the histogram method or whole signal cross-correlation using equation (3-10). In both cases the method is applied after the input filters. A version of PSQM was also produced, using the same delay estimation processes operating after the input filters. In addition, the processing for the histogram method was also performed in PESQ with pre-filtering removed; this is equivalent to using $\psi(\Omega)=1$ in Knapp and Carter's generalised structure.

The results of this comparison are summarised in Table 3.3, which shows the mean and worst-case correlation for these methods, applied to the estimation of a single delay for the whole signals, in the constant-delay subjective tests from Appendix D. The highest average performance for each model is found using the histogram method rather than signal cross-correlation: for PESQ, the histogram method (1) gives 0.7% higher average correlation than for cross-correlation (2); with PSQM, the difference between the methods ((4) and (5)) is 0.34%.

Table 3.3: Comparison of constant-delay estimation algorithms

Model	Mean correlation	Worst-case correlation
(1) PESQ, histogram method, whole signal	0.9532	0.8828
(2) PESQ, cross-correlation, whole signal	0.9462	0.8838
(3) PESQ, histogram method, whole signal, no pre-filtering	0.9533	0.8846
(4) PSQM, histogram method, whole signal	0.8199	0.4677
(5) PSQM, cross-correlation, whole signal	0.8165	0.4633

Two results in Table 3.3 are slightly surprising: whole-signal cross-correlation (2), and the histogram method applied without pre-filtering (3), both give better worst-case performance than the standard method (1). This is investigated further in Figure 3.18, which shows the distribution of the difference in correlation coefficient between the modified methods and the standard histogram method with pre-filtering, for the constant-delay subjective tests.

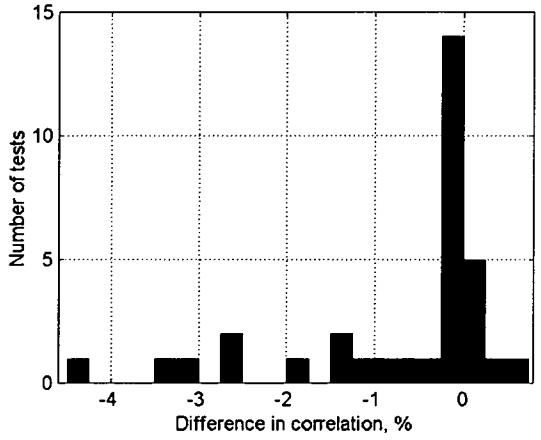
For 9 of the tests, the cross-correlation method gives correlation that is more than 1% lower than the standard histogram method, as shown in Figure 3.18(a); in the most extreme case, cross-correlation reduces the model's correlation coefficient by 4.3% compared to the histogram method. However, in 7 of the tests, cross-correlation gives slightly better accuracy, up to 0.65% higher in one case. This was found to be due to cases of temporal clipping leading the crude alignment to give estimation errors outside the range of capture of the histogram alignment, a problem avoided by cross-correlating the whole signals. However, the variable-delay alignment processes described in the next section corrects many of these mistakes.

Removing pre-filtering from the delay estimation, plotted in Figure 3.18(b), gives a roughly even split between tests that are improved and those that are made worse. This indicates that,

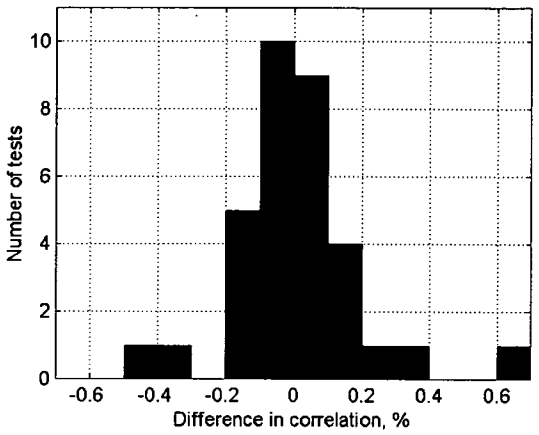
for the narrowband signals used in telephony, pre-filtering has limited benefit, and suggests that the degree of bias shown in Figure 3.7 is likely to be rare within the scope of this thesis.

Figure 3.18: Comparison of delay estimation algorithm performance with PESQ

(a) Histogram method compared to cross-correlation



(b) Histogram method with pre-filtering compared to histogram method without



3.5 Robust identification of variable delay

The algorithm described in the previous section provides an improvement over existing methods, but it is still constrained to estimation of a constant delay. Early on in the work described in this thesis it became clear that the assumption of constant delay throughout a measurement would lead to significant inaccuracy in certain cases. This part of the chapter describes the problem in more detail and outlines the two stages of solution. In the first stage, constant delay is identified using the method described in section 3.4 for each utterance, typically containing 1–3s of speech. This provides reasonable control of estimation error – which may be high with shorter sections of signals, a problem with the dynamic time warping (DTW) method used by Beerends and others – but is able to deal with the most common form of delay variations in VoIP, during silent periods. The second stage deals with cases where the first stage fails because an incorrect estimate is made or the delay changes during an utterance, by splitting utterances and realigning each part, accepting the best candidate in terms of delay confidence using a maximum likelihood approach.

The first of these methods, the alignment of separate utterances, was developed in conjunction with Reynolds, who conducted the first explorations into "Segmental PAMS" and collected the data described in section 3.5.1.2. The author implemented this method in PAMS 2, which was the first perceptual model designed for end-to-end testing of networks including VoIP. The second method, utterance splitting, was based on the author's own investigations and those of Beamond, who was working under his direction. After improving the way delay

changepoints were identified and making the algorithm more computationally efficient, the author incorporated this into PAMS 3, which was entered in the ITU-T competition that led to P.862. The time alignment routines from PAMS 3 were copied in PESQ [ITU-T P.862, Rix 2000c, Rix 2002b], replacing the DTW algorithm used in PSQM99. These methods were the subject of a successful patent application [Rix 1998c].

3.5.1 Types of delay variation

Delay variation in speech communications systems may arise from many sources. Of these, the most common is packet-based transmission such as VoIP. These sources are considered in two categories based on the result in the measured audio signals: continuously variable delay and piecewise constant delay.

3.5.1.1 Continuously variable delay

The problem of continuous delay variation for perceptual quality assessment pre-dates this work, and was considered in the late 1980s. The application in this case was assessment of high-quality audio in real-time streaming, particularly with analogue magnetic tape. Here variations in the tape transport speed result in continuous delay changes, which were addressed by dynamic time-warping (see below). This was reportedly successful [Herre 2001]. However, in this application group delay distortion is likely to have been minimal and the waveform coders being assessed are of much higher quality than speech coders, so the results do not necessarily apply to speech communications.

Similar continuous delay variation may be caused by clock skew or sample rate jitter. Even with transport systems such as synchronous digital hierarchy (SDH), which are mainly based on optical transmission, the lower-order digital trunks commonly run asynchronous clocks, although generally with tolerances on the order of 0.01% or better. Sample rate variations in measurement equipment are of a similar order. In theory this could cause up to 1ms skew over a 10s measurement. This is large enough to cause serious problems with linear methods such as those outlined in sections 3.3 and (for transfer function estimation) section 4.3, but has only a small effect on a perceptual quality measure such as PAMS or PESQ as long as it remains a small fraction of the temporal resolution of the auditory transform.

Analogue frequency division multiplexing (FDM) transmission systems using single side-band (SSB) modulation may lead to a slightly more complex effect, where discrepancies in the carrier between the transmitter and receiver lead to continuous phase variation between the two ends. This results in non-harmonic frequency shifting, which was found in informal listening tests to be perceptible above about 0.1%. However, tolerances on these systems are normally much

better than this, and analogue FDM has now been almost completely replaced by photonic transmission systems such as SDH in most Western markets.

Changes in acoustic and/or radio path may also cause continuous delay variation, but this is on a much smaller scale. The acoustic path, if present, is normally held constant during a measurement. Tests of a live analogue mobile network may lead to delay changes, but since a speed of 100kph is approximately 10^{-7} of the speed of light these are unlikely to be significant for our application.

In summary, the most serious source of continuous delay variation for communications applications is clock variation between transmit and receive devices. The author has found that continuous delay variation is adequately dealt with by a piecewise constant delay estimation algorithm, provided it has sufficient granularity: 0.01% clock difference results in a variation in time-delay between the signals of 0.2ms over 2s, which is a typical duration of a spoken utterance. Figure 3.1 shows that this is likely to have minimal effect on a perceptual model.

3.5.1.2 Step delay variations

In contrast to this, several different types of system may cause large step changes in delay. Prior to the work of the author and of Reynolds [Rix 1998c, Rix 1999b], the consequences of this phenomenon on perceptual models had not been described in the literature.

The author first encountered delay variation with digital circuit multiplication equipment (DCME) that was able to switch bit-rate. Measurements were being performed using a script that included an echo canceller disabling tone prior to the speech recordings. Echo-canceller disabling is normally performed by modems and faxes, and to ensure transparency with these the DCME switched to its highest rate (40kbit/s ADPCM). However after a period of natural speech – which may be as low as a few seconds, but in this case was 20–30s – the DCME reverted to speech transmission at 16kbit/s using a CELP codec. This resulted in an end-to-end delay increase on the order of 10ms, which was enough to cause a large drop in PAMS score [Rix 1998b].

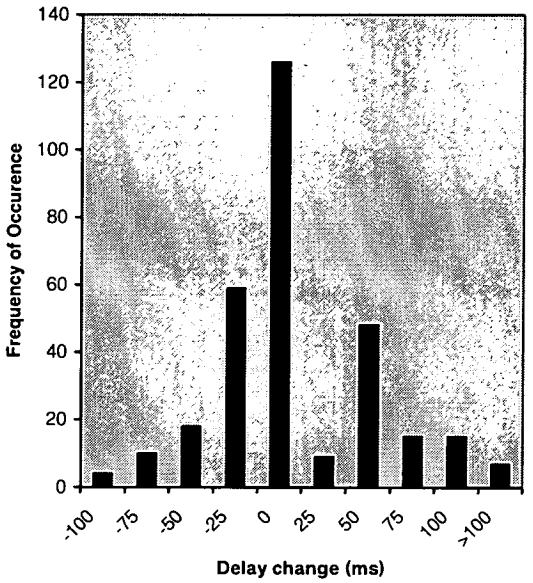
Bit-rate variation to take account of traffic conditions is becoming more common, but the DCME case was unusual. Echo-canceller disabling is only intended for data traffic and should never normally be used with speech; the script in this case was badly designed. Modern systems that may exhibit bit-rate variation use scalable codecs such as AMR [GSM 06.90] or EVRC [TIA/EIA IS-127]. In these cases the framing, and hence delay, is kept constant; the only changes are in the number of bits of payload transmitted, and the error control coding used for each frame.

VoIP is now by far the most common source of delay changes. This is clearly shown by the data gathered by Reynolds, which was reported in [Rix 1999b]. This is shown in Figure 3.19. In

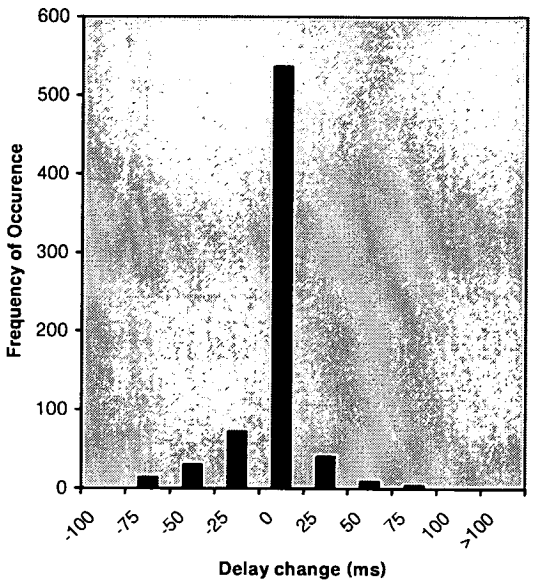
the trial of voice over modem-based IP dial-up (Figure 3.19(a)), more than 50% of conditions showed a change to the end-to-end audio delay of 25ms or greater during the 16s measurement. Changes of this magnitude are sometimes perceptible, and caused severe problems with previous perceptual models such as PSQM [ITU-T P.861] and PAMS 1.

Figure 3.19: Delay variation in VoIP

(a) PC-PSTN VoIP field trial



(b) VoIP gateway evaluation



3.5.1.3 Origins of delay variation in VoIP

It is becoming increasingly common to use packet-switched networks, typically based on Internet protocol (IP) and/or asynchronous transfer mode (ATM), for carrying real-time speech traffic. This has the potential for significant cost savings over traditional circuit-switched networks, due to the lower cost of switching equipment and the ability to run a single network for both voice and data. For the purposes of this thesis, the main focus is voice over IP (VoIP).

In packet-based transmission, speech is compressed using a coder such as G.711 or G.723.1, and divided into packets. In VoIP a typical packet length is 20ms, a trade-off between delay and increased bit-rate due to the packet headers [Reynolds 2001a, Reynolds 2001b]. The packets are sent across the network, reassembled and decoded to a speech stream at the receiver. Packets may reach the receiver in a different order, because the route used to transmit the packets may change to one of different delay. Additionally, because many network components operate on a "best effort" basis and switch traffic from many streams, the time taken for each packet to travel may vary and some packets may be lost.

The variation in the delay of each packet is termed packet jitter. It is not usually a problem for data traffic, which is fairly insensitive to delay, allowing time to request that lost packets be re-sent. Re-sending is not normally feasible for two-way voice communication, which can be hampered by round-trip delay of as little as 50ms if there is echo present. Although the impairment due to delay can be reduced by use of appropriate echo cancellation, this cannot eliminate the effect of delay on conversational quality.

There are therefore two conflicting requirements. To prevent packet jitter from leading to packets being discarded because they arrive too late, the receiver must buffer the incoming data. The longer the buffer, the fewer packets are lost and the higher the (one-way) speech quality of the system. However, a longer buffer means greater round-trip delay, and correspondingly lower conversational quality. From a conversational perspective, the buffer needs to be as short as possible.

Packet jitter in the network can lead to variations in the delay in the audio path through several different mechanisms. Two of the most common are dynamic buffer resizing and excessive late packet dropping.

Dynamic buffer resizing during silence is a common method used in VoIP systems to deal with time-varying levels of packet jitter whilst attempting to minimise bulk delay. The buffer length is changed during silent intervals and leads to delay variations in silence.

Excessive late packet dropping is a less frequent effect. In this case a large change occurs in the packet delay during a speech event – for example, a route alteration that delays subsequent packets by 100ms more than the previous route. Following this change the first few packets arrive too late and the buffer becomes empty. The typical result is the insertion of silence or interpolated concealment frames in the middle of the speech event, with playback re-starting only after a few new packets have arrived in the buffer.

3.5.2 Dynamic time-warping

A technique developed for template-based speech recognition, dynamic time-warping (DTW), performs frame-by-frame delay adaptation using cross-correlation prior to, or as part of, feature extraction and matching. This has been used by other authors for time-varying delay identification in perceptual models.

This is suited to speech recognition, where duration and pitch vary significantly between instances of the same utterance, even for the same talker. The term DTW is usually applied to time alignment in feature space using dynamic programming methods [Silverman 1990, Keogh 2001]. A simpler frame-by-frame DTW method implements delay adaptation using a greedy approach [Cormen 1990] rather than dynamic programming, and has been applied to perceptual models by Herre and others for audio to compensate for tape playback rate variation

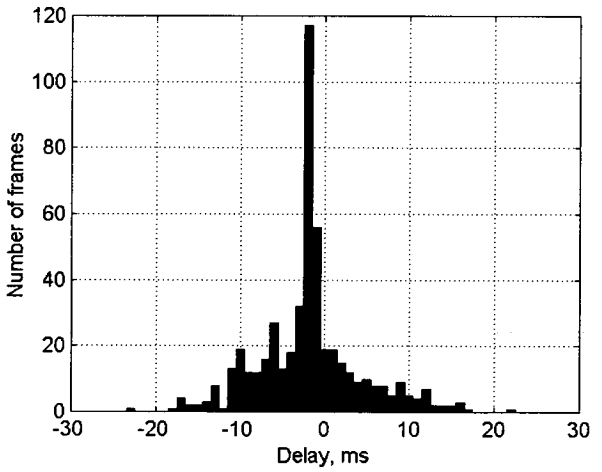
[Herre 2001], and by Beerends and Hekstra in PSQM99 for speech quality. This is the method considered in this section.

DTW performs windowed frame-by-frame cross-correlation, using the method shown in equations (3-11) and (3-12), to compute a delay estimate $\hat{\tau}(k)$ using cross-correlation for frame k. It is assumed that crude delay has been estimated using a method such as that described in section 3.4.1.2, and eliminated from the calculation. However, unlike the methods described in section 3.4, no inter-frame smoothing is performed; the delay estimate $\hat{\tau}(k)$ is used directly for the processing of the corresponding frame in the perceptual model.

If distortions are small, DTW should be able to measure both continuous and step delay changes provided that they are much less than the frame size N_f. However, group delay variation combined with time-varying spectral content of the signals, or heavy distortion, mean that a significant proportion of frames may be misaligned. This can be seen in the histogram of DTW delay estimates shown in Figure 3.20, which is based on the example constant-delay condition introduced in Figure 3.12 and Figure 3.13. Bulk delay has been removed in this example using the crude delay estimate calculated by the algorithm presented in section 3.4.1. The signal pre-filtering from PESQ that is described in the same section was performed on both reference and degraded signals, to reduce the bias effect discussed in 3.3.3.2.

Figure 3.20 shows substantial scatter in the delay estimates for each frame. This may cause the distortion measures to reduce or increase, depending on whether the DTW delay finds a better match – a more similar waveform – than at the "true" delay, or is made incorrect by the presence of loud distortions or non-linearity. For this example, the change in PESQ score is relatively small: 2.585 with P.862 PESQ; 2.565 with DTW using 64ms frames using a Hann window. Further results on performance with DTW are given in section 3.6.

Figure 3.20: Distribution of DTW delay estimates



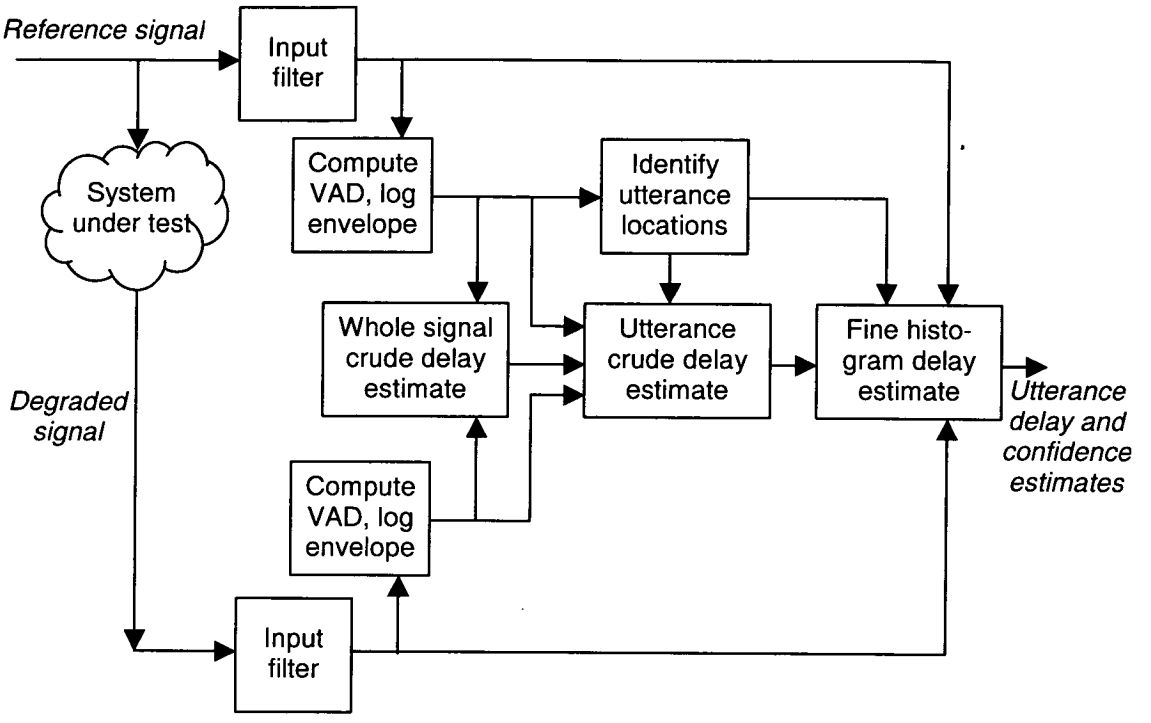
A further problem with DTW is that delay changes may in practice be too large for the DTW method to locate. A typical frame size used for DTW is 64ms, allowing delay changes up to about ±20ms to be identified. However Figure 3.19 shows that delay changes as large as 100ms may occur, and these are likely to lead to unstable and inaccurate estimates $\hat{\tau}(k)$ for this frame size. Whilst raising the frame size reduces this problem, it makes DTW more susceptible to certain types of non-linearity, and increases computation time.

3.5.3 Utterance delay estimation

If it can be assumed (a) that accurate time-delay estimation is improved by considering sections that are as long as possible, and (b) that the most common delay changes are step delay changes during silent intervals between speech utterances, the most accurate method is likely to be to estimate the delay of entire speech utterances. This was found to be preferable to DTW when the assumption that delay is constant throughout the utterance is correct; rather than using individual frames for each delay estimate, a single estimate is calculated using information from all frames in the utterance.

The concept for this algorithm was first suggested during discussions between the author, Reynolds and Gray. Reynolds produced an early implementation in Labview that gave a similar effect, by invoking PAMS separately for each utterance and aggregating the distortion parameters. The author improved the method by adding padding and performing an initial crude delay estimation to eliminate large constant delays, and integrated it into the perceptual model. The improved algorithm is shown in Figure 3.21. C code for this method is given in functions utterance_locate(), id_searchwindows() and id_utterances() in pesqmod.c in [ITU-T P.862].

Figure 3.21: Algorithm for utterance delay estimation



The process begins by computing the VAD and log envelope described in section 3.4.1.2. This is used to compute an initial crude delay estimate over the whole signal duration, to cancel out large bulk delay from the subsequent processing. Speech utterances are defined from the VAD as continuous portions of speech of at least 300ms duration; this minimum duration limit is necessary because it is difficult to compute a reliable delay estimate in coding distortion for shorter sections. The identified utterances are realigned using the same crude delay estimation method (section 3.4.1.2) to eliminate large delay variations. Finally, the histogram-based fine delay estimation algorithm described in section 3.4.2 is used to compute delay and delay confidence measures for each utterance.

The first crude delay estimate takes account of the bulk delay of the system. In the event of asynchrony between measurement devices, this can often be longer than 1s, while utterances can be as short as 300ms. The second, utterance-based, crude delay estimate is also required because, as illustrated in Figure 3.19, delay variations may be much larger than the capture range of the fine delay estimation process.

Note that a limitation of this method is that it does not localise delay changepoints during silent intervals. It was found to be very difficult to perform this in a general way because in many cases DTX, which is now common, destroys any relationship between the waveforms during comfort noise periods; however, for clean speech reference signals, the location of delay changepoints during silent intervals makes little difference to the quality score. The placing of

utterance start and end points for perceptual modelling of delay changes during speech is considered in section 3.5.7.

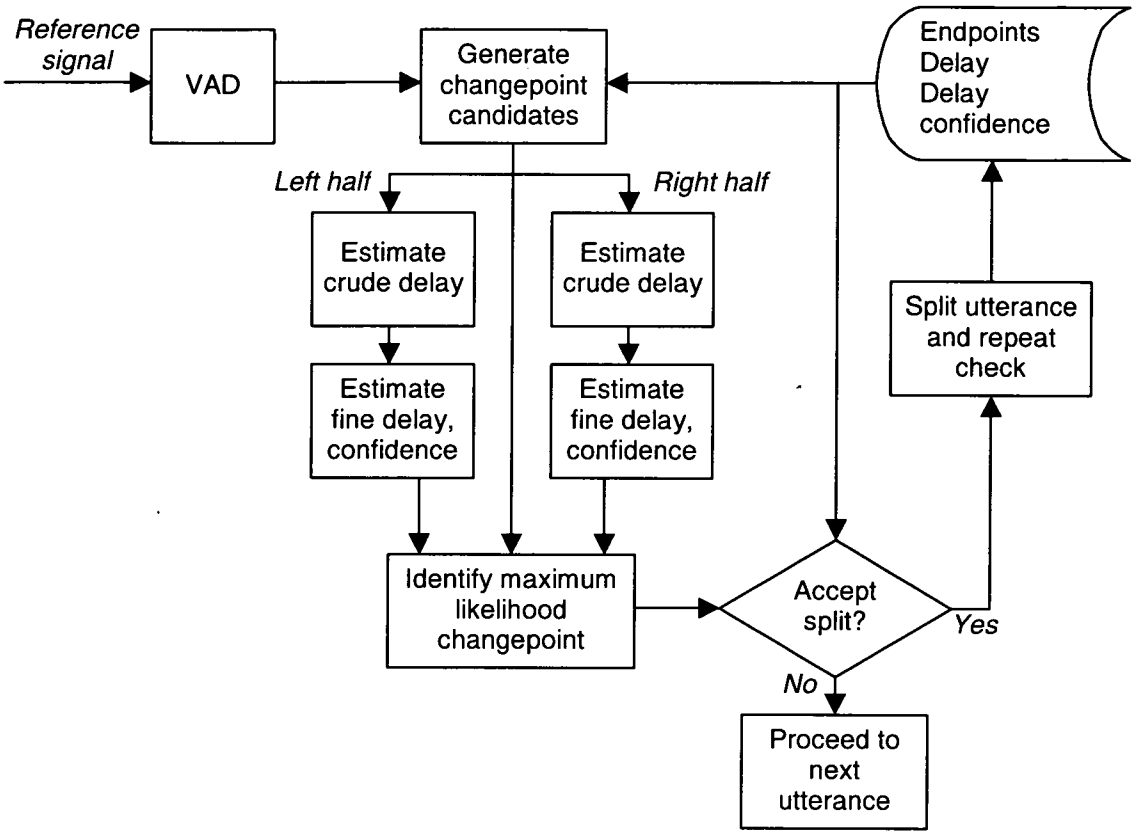
3.5.4 Utterance splitting

While delay variations most commonly occur during silent intervals, it is also possible for delay to change during a speech utterance, for example due to re-routing in a packet-based system. The utterance alignment process, while accurate in many cases, is not able to identify these delay changes and normally identifies the delay that applies to the longest part of each utterance. The sections of each utterance that are incorrectly aligned lead to large measured errors, giving a quality score that is too low.

To address this problem while using sections that are as long as possible, the author extended the principle of utterance time alignment to the recursive splitting and realignment of utterances to localise delay changes, and developed an efficient method to achieve this. Some investigative work on utterance splitting was carried out by Beamond, who worked for the author during a summer placement and considered performing a golden section search to identify optimum changepoints [Rix 1998c]; this method was found to be less efficient than the method described below.

The general algorithm for utterance splitting to identify a delay change is set out in Figure 3.22. This considers splitting each utterance in turn at up to 40 locations; a minimum duration of 300ms for each split part is maintained to reduce the likelihood of misalignment. For each split, the halves of the utterance before and after the split are separately realigned by crude alignment followed by fine alignment using the algorithms of sections 3.4.1.2 and 3.4.2, and the confidence is calculated as described in section 3.4.2.2. If a split identifies a delay change and results in an increase in delay confidence for both split halves, the best such delay change is used to divide this utterance in two, and the entire process is repeated to search for further delay changes in each of the new halves; otherwise, it continues to the next utterance.

Figure 3.22: Utterance splitting algorithm



A heuristic approach was followed in the development of PAMS for the identification and acceptance of the most likely changepoints. A number of different minimum thresholds for the delay change were tested, and the value of 4ms was found to provide sufficient accuracy while being sufficiently high to prevent excessive numbers of false splits being detected in constant delay conditions. Several alternative criteria were evaluated for comparing the delay confidence estimates to identify the "best" candidate and to determine whether to accept the split; no significant advantage was found in using more complex tests such as non-linear averaging, the product of the confidence values, or weighting compared to the following simple criteria.

- The best changepoint has the highest sum of delay confidence for the left and right halves.
- The best changepoint is accepted if the delay confidence for each half exceeds the delay confidence of the whole utterance.

Section 3.4.3 developed a MAP approach to delay estimation, in which the histogram is replaced by the log likelihood function. Using this framework, the utterance split decision can be considered as a comparison between a model where the utterance is undivided, and a model with a changepoint and the candidate delays of each half. (Potentially further splits could also be considered, although this further increases the computational complexity of the problem.)

Bayes' theorem (3-17) can be used to incorporate prior knowledge about the likelihood of delay changes during speech and formulate the splitting process as a MAP model selection problem.

3.5.5 Improvement of utterance splitting

The algorithm shown in Figure 3.22 is computationally intensive because, even for constant delay conditions, both parts of each utterance are realigned for every changepoint that is considered – up to 40 times. This was found to cause the time alignment in PAMS to take more than twice as long as all other model components combined. The author therefore improved the method to avoid repeated computation.

The main complexity of the algorithm of Figure 3.22 is in the FFTs used to evaluate the cross-correlation of the envelope and individual frames. For envelope cross-correlation an exact method is used with zero padding to eliminate circular effects; for the time alignment frames the window minimises circular effects and no zero padding is used. The following example considers the main processing required in splitting a 2s utterance in the middle, at 8kHz sampling rate, using 64ms 75% overlapping frames for the histogram alignment. The envelope frame size is 4ms. (Θ notation is not used here because the constant multipliers are of interest.)

- Crude alignment by envelope cross-correlation for 1s sub-utterance (250 envelope frames) using efficient FFT-based method: 2 sub-utterances x 3x 512-point real FFTs.
- Windowed histogram: 2 sub-utterances x 62 alignment frames x 3x 512-point real FFTs.

In this example, the evaluation of the windowed histogram takes more than 20 times the computation of the envelope correlation. However, if the crude delay is found to be constant for a number of candidate changepoints, the windowed histogram can be assembled progressively, requiring cross-correlation to be performed only for new frames. Normalisation and kernel smoothing to estimate the delay and confidence are carried out separately, without overwriting the current histogram.

For 40 changepoints the improved method requires just over 1kB of additional storage. In the best case of a constant-delay condition where all crude alignments produce the same delay estimate, the number of FFT evaluations for the windowed histogram is reduced by a factor of 40; in the worst case, the complexity will be the same as the algorithm of Figure 3.22. C code for this improved method is given in the functions utterance_split() in pesqmod.c, and split_align() in pesqdsp.c in [ITU-T P.862].

3.5.6 Realignment

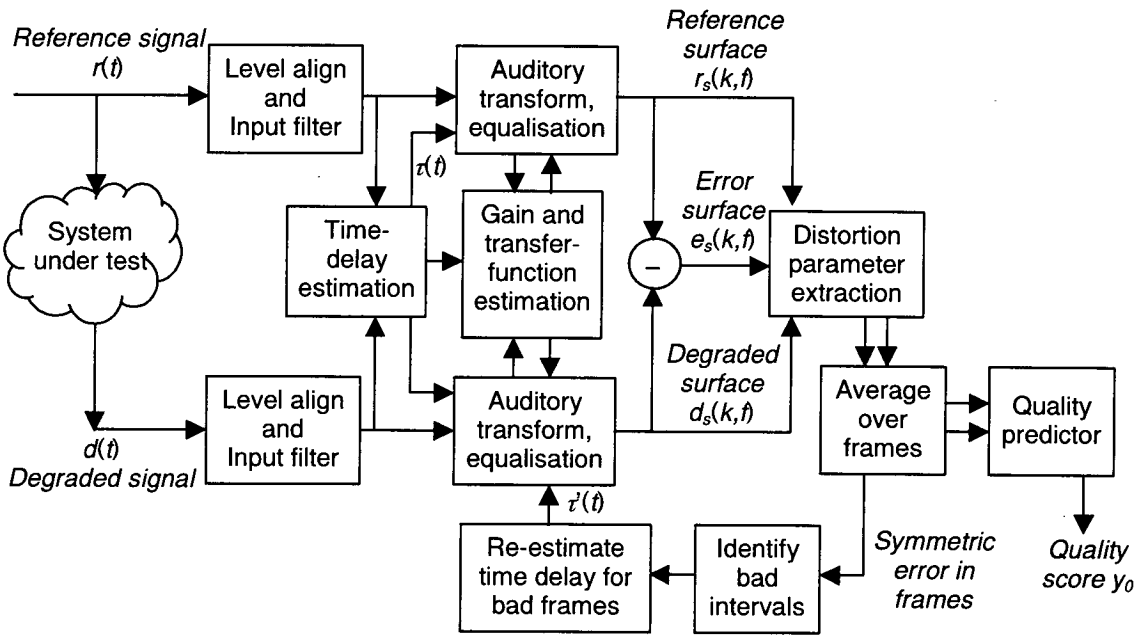
For PESQ, Hekstra developed a further improvement to time alignment [ITU-T P.862, Beerends 2002]. The processes of utterance alignment and splitting may fail to identify some delay

changes, particularly if the delay change occurs close to the start or end of an utterance, or lasts for less than about 200ms. A symptom of these failures is that very large distortions are observed.

Hekstra's method proceeds as follows. Firstly, the main part of the perceptual model, including time-delay estimation, is executed and the distortion parameters are averaged over frequency to compute a measure for each frame. If the symmetric distortion parameter in any frame exceeds a pre-determined threshold, a bad interval is identified. Bad intervals are extended in each direction in time by two frames (32ms) to provide some padding at start and end and to join together intervals separated by fewer than four good frames.

Cross-correlation of the reference and degraded signals for each bad frame is used to estimate the error in time-delay for the bad interval. The auditory transform, including gain equalisation, is re-run for the corrected delay. If the result is a decrease in either distortion parameter for any frame, the lower value is used. Once all bad intervals have been realigned no further iterations are performed; the distortion parameters are then averaged over time and the overall PESQ score is computed. This process is shown in Figure 3.23.

Figure 3.23: Bad frame realignment in PESQ



In PESQ, the processing to implement one iteration of bad interval re-alignment is performed in function pesq_psychoacoustic_model() in pesqmod.c in [ITU-T P.862].

3.5.7 Perceptual modelling of delay variations

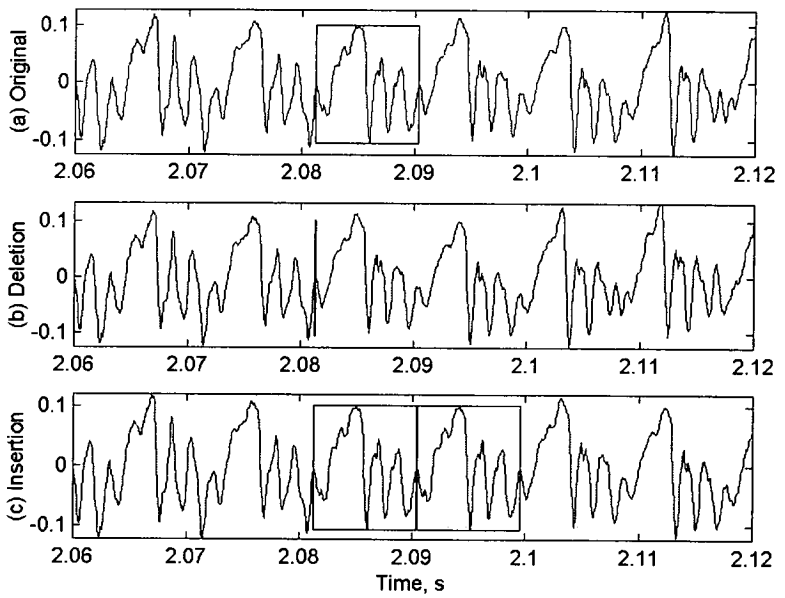
The previous sections have presented methods which estimate the delay during a measurement and are able to localise delay changes to within a small fraction of the duration of an utterance. This section discusses how this information is used. Before this work, no other author had published on how delay variations can be included in a perceptual model.

The starting point for incorporating time-varying delay into a perceptual model is the audibility of any resultant distortions. This is determined by whether a perceptible discontinuity is introduced. Delay changes during silent intervals are likely to be imperceptible as long as no transient is introduced. It is also possible to make significant delay changes to a waveform during speech without causing an audible distortion, provided that:

- there is no waveform discontinuity; and
- the signal pitch and formant structure is preserved, for harmonic signals such as voiced speech or music.

For example, during voiced speech, an entire pitch cycle may be deleted or inserted with little perceptible effect. This is illustrated by the example male voiced syllable /a/ shown in Figure 3.24. The 9.2ms waveform section enclosed by the box in Figure 3.24(a) is removed in Figure 3.24(b) and duplicated in Figure 3.24(c). Informal listening to this example shows no perceptible distortion.

Figure 3.24: Delay changes during voiced speech



In practice delay changes during speech are seldom concealed as effectively as shown in Figure 3.24. In VoIP, packet sizes of 20–30ms are common for low bit-rate coders [Reynolds 2001a]. Speech parameters such as pitch may change significantly over these periods; a packet may even contain an entire acoustic event such as a stop consonant. In these cases deletion or insertion of some concealment signal is likely to cause some audible distortion.

The author addressed this issue in PAMS 3 by treating deletion and insertion separately, as follows [Rix 2000b]. Where deletion occurs (i.e. a section has been removed in the degraded signal), the reference signal is processed continuously; the section of the degraded signal in the vicinity of the deletion is therefore processed twice. Conversely, if insertion occurs, the degraded signal is processed continuously and the section of the reference at the location of the insertion is processed twice.

By sampling across the delay change in either case, any audible discontinuity or poor concealment will be measured by the perceptual model. If concealment has been performed well and the signals are near-stationary in the region of the changepoint, as in Figure 3.24, no distortion will be observed. This operation is implemented when the utterance start and end positions are set during the utterance splitting process, in utterance_split() in pesqmod.c in [ITU-T P.862].

Due to pressure of time during the integration of PAMS and PSQM99 to produce PESQ, this process was not included in PESQ. An alternative solution by Hekstra and Beerends is used instead. In PESQ, processing is always aligned with the reference signal and the utterance boundaries are used only in converting delay to this time-base. In the event of insertion, this means that a section of the degraded signal is not processed; if this coincides with the inserted section, the model cannot tell the difference between good and bad concealment. In the event of deletion, a section in the degraded signal is processed twice, but the disturbance is set to zero in this region if the delay change is greater than the frame period (16ms) [ITU-T P.862]. Thus where PAMS specifically processes across delay changes, PESQ specifically ignores them. This accounts for the greater accuracy of PAMS for the VoIP subjective tests that involve delay changes during speech.

The author has implemented a fix for PESQ to include the same processing over delay changes as in PAMS. The performance of this corrected model (number (7)) is presented in the next section.

3.6 Results

This section presents results to illustrate the performance of perceptual models that include the histogram-based alignment and the variable delay identification techniques developed in this chapter. A more detailed analysis of the constant-delay estimation algorithms developed in

section 3.4 was presented in section 3.4.4; only the results for PSQM and PESQ with constant-delay alignment using the histogram method are duplicated from that section.

The following models are compared.

(1) PSQM [ITU-T P.861] extended with constant-delay alignment using the histogram method.

(2) PAMS 3.1 (January 2001), using the histogram method with utterance splitting and processing over delay changepoints.

(3) PESQ with only constant-delay alignment using the histogram method.

(4) PESQ with only crude alignment followed by DTW.

(5) PESQ with separate utterance alignment but without utterance splitting or Hekstra's realignment.

(6) PESQ with utterance splitting but without Hekstra's realignment.

(7) PESQ extended with processing over delay changepoints.

(8) PESQ as standardised in [ITU-T P.862].

The performance of these models over the subjective test database described in Appendix D is given in Table 3.4. This presents mean and worst-case correlation coefficient, calculated as described in section 2.6, for the 33 constant-delay subjective tests, for the 12 variable-delay tests, and for the whole set of 45 tests.

Table 3.4: Time-delay estimation and perceptual model performance

	Constant-delay tests (33)		Variable-delay tests (12)		All subjective tests (45)	
Model	**Mean correl.**	**Worst-case correl.**	**Mean correl.**	**Worst-case correl.**	**Mean correl.**	**Worst-case correl.**
(1) PSQM, ITU-T P.861	0.8199	0.4677	0.6608	0.2792	0.7775	0.2792
(2) PAMS 3.1	0.9423	0.8030	0.8946	0.7645	0.9296	0.7645
(3) PESQ, histogram method, whole signal	0.9532	0.8828	0.7678	0.5271	0.9038	0.5271
(4) PESQ, DTW method	0.9533	0.8856	0.8022	0.6017	0.9130	0.6017
(5) PESQ, no splitting or realignment	0.9542	0.8994	0.8714	0.6528	0.9321	0.6528
(6) PESQ, no realignment	0.9547	0.8994	0.9109	0.8055	0.9430	0.8055
(7) PESQ, with temporal discontinuity model	0.9551	0.9000	0.9121	0.8063	0.9436	0.8063
(8) PESQ, ITU-T P.862	0.9550	0.9007	0.9118	0.8108	0.9435	0.8108

PSQM (1) is the weakest of these models. Even for constant-delay subjective tests, its worst-case performance is poor, as a consequence of the problems with linear filtering and quality prediction that are discussed in the following chapters.

PAMS (2) performs much better than PSQM. Both for the variable-delay tests and for the overall dataset, it is also more accurate than PESQ with DTW (4), which is most similar to the PSQM99 model entered in the ITU-T P.862 competition. However, with inclusion of the variable delay routines from PAMS, PESQ (6) is more accurate than PAMS both on average and in worst-case performance.

It is likely that the constant-delay subjective tests include some conditions with delay variations, because all of the variable-delay versions of PESQ (models 4–8) perform better than the constant-delay version (3) on the constant-delay dataset. As described in section 3.5.1, clock jitter is a likely cause of this difference in many cases. This also indicates that providing more degrees of freedom for the variable-delay algorithms does not degrade overall perceptual model performance.

The difference between models (3) and (5) is the utterance alignment technique presented in section 3.5.3. This clearly provides a substantial improvement in average and worst-case correlation for variable-delay subjective tests.

Comparing models (5) and (6) shows that the utterance splitting process described in section 3.5.4, which is included in (6), also provides an improvement, particularly in worst-case performance for subjective tests that include delay changes during speech.

The remaining improvements described in sections 3.5.6 and 3.5.7 provide limited further gains. Hekstra's bad frame realignment gives P.862 PESQ (8) about 0.5% better worst-case performance, and 0.1% improvement in average correlation for the variable delay tests, compared to the model without realignment (6). The temporal discontinuity processing from PAMS gives a fractional improvement in average performance in model (7), but reduces worst-case correlation. This is thought to be due to interactions with the muting problem that is discussed in section 2.5.4.

The value of the variable delay identification techniques introduced in the previous section is clearly shown by comparing P.862 PESQ (8) with PSQM (1), the constant-delay version of PESQ (3), or the DTW method (4). On the variable-delay dataset, the worst-case correlation of model (8) is higher than the mean correlation of models (1), (3) and (4). Further tests of the differences between models are given in section 5.8.

3.7 Conclusions

The cross-correlation method for time-delay estimation may lead to inaccurate results with systems that include significant dispersion or non-linearity. The method of constructing a smoothed histogram of delay estimates appears to give a delay estimate that is more robust both to noise and to coding distortions, giving a substantial improvement in perceptual model performance in some cases. In addition, the new histogram method provides a confidence measure that can be used to identify delay changes.

Time-varying delay is a particularly important issue for VoIP, and must be taken into account for a perceptual model to provide accurate scores with these technologies. The methods of utterance alignment and utterance splitting provide significant improvement in perceptual model performance for VoIP conditions, with minimal loss of performance for constant-delay tests, and are much more effective than the DTW method. The perceptual realignment process makes a small additional improvement in accuracy for variable-delay tests. These techniques mean that, unlike PSQM or other early perceptual models, it is possible to use PESQ and PAMS 3.1 to give accurate quality measurements for telephone networks where delay changes may occur.

A difference between PAMS and PESQ is the way that delay variations are processed in the auditory transform. However, incorporating the PAMS method of sampling across delay changepoints into PESQ was found to make little difference in accuracy for tests in which these conditions are common.

Even with these innovations, PESQ is not quite as consistently accurate for variable-delay subjective tests as it is for constant-delay tests. In part this may be because the variable-delay tests are more critical, including a wider variety of network conditions, but the muting problem may contribute to this. In particular, the worst-case correlation for the variable-delay tests, 81%, indicates that there is room for improvement.

Areas for further work include studying the localisation accuracy of the utterance splitting process, investigating and implementing the delay identification and utterance splitting algorithms in a Bayesian framework, and enhancing the performance of PESQ in the most critical variable-delay conditions.

CHAPTER 4

Transfer function equalisation

4.1 Overview

The strength of the method of comparison of auditory transforms is that it allows the perceptual model to estimate the audibility and loudness of distortions. This works well for the assessment of speech and audio coders, where the interface is digital, as the envelope and signal spectrum are largely preserved and the main errors are non-linear coding distortions. However, linear filtering with a non-flat frequency response, such as that found in analogue links of telephone networks, results in large systematic shifts in the auditory transform and leads to substantial errors being measured. Filtering is everywhere in the acoustic environment that we use for speech communication from day to day; the human ear adapts automatically to strong filters such as room echoes or the person we are listening to turning their head away. These are not perceived as distortions. The result is that perceptual models such as PSQM that do not distinguish linear and non-linear distortions produce scores that are too low in the presence of linear filtering. Example results in the next section show that this severely limits their accuracy.

The limited subjective effect of "spectral tilt" was noted by Quackenbush [Quackenbush 1988], who suggested that the bulk of the effect could be eliminated not by frequency response equalisation but by frame-by-frame equalisation of overall amplitude. A similar process to this is included in PSQM, and in practice does not provide sufficient compensation for linear filtering.

At the start of the work described in this thesis, perceptual models for audio quality measurement were beginning to include a frequency response adaptation process [Thiede 1996]. In audio coding, the main linear distortions are bandlimiting as part of the encoding process, or small amounts of linear filtering associated with high-quality analogue audio connections. Bandlimiting is used by most audio coders at low bit-rates, where there is not enough capacity to encode the high frequencies adequately and the bits are better spent providing better resolution at low frequencies. The result of bandlimiting is that there is no information in the degraded signal above the cut-off frequency, and significant (tens of dB) anti-aliasing filtering just below this. Analogue interconnection is normally much less severe, but anti-aliasing filters or component mismatch can still cause several dB of ripple or filtering in the

main signal band, and this is still large compared to audio coding distortion. Bandlimiting is quite audible and produces some noticeable reduction in perceived quality, but the filtering due to analogue interconnection is normally inaudible.

Thiede developed a pattern adaptation process to address both of these effects as part of a perceptual model for audio quality assessment [Thiede 1996]. The process was subsequently included in PEAQ [ITU-R BS.1387]. This operates by adaptively filtering the output of each perceptual band of the reference signal to equalise it to the degraded signal, using smoothing in time and frequency to derive the required gains. This method is able to adapt to both slow gain and transfer function variations, which may also be relatively inaudible.

This dynamic approach to frequency response equalisation is not necessarily optimum for speech quality assessment. The possibility for high additive noise, and the non-linearity of coders and algorithms such as noise suppressors means that large variations in the short-term frequency response can occur, which may either be masked by a dynamic equalisation or cause errors to be over-estimated if the dynamic adaptation overshoots. For these reasons, Thiede's dynamic method is not considered further in this text. These properties also mean that conventional techniques for stationary linear system identification are of limited use because they may give biased results, as shown in section 4.3.

During this work four different approaches to transfer function equalisation have been published in the context of perceptual speech quality measurement. The author developed a method for phaseless transfer function estimation, operating in the perceptual filterbank domain, which is presented in section 4.4.3 and was implemented in PAMS [Rix 1999b, Rix 1999f]. Berger used the estimated total frequency response to equalise the reference signal in TOSQA prior to the auditory transform, but little detail has been published on how this is achieved [Berger 1997]. Beerends and Hekstra calculated the bark spectra over the whole signals and used the spectral difference to equalise the reference to the degraded; this method was implemented in PESQ [ITU-T P.862; Beerends 2002]. Finally, Park modified BSD to compute errors using a coherence function that automatically eliminates the gross frequency response [Park 2000]. These approaches are described in section 4.4, and results on the effect of transfer function equalisation in PESQ are given in section 4.5.

4.2 Linear filtering in communications networks

4.2.1 Components that introduce filtering

Many components of a telephone network may filter the signal. These include the following.

Acoustic path. The physical shape of the telephone device, human head and torso, and their relative positions, along with the response of the room, all behave as linear filters. Other than the close-coupled response between the mouth and handset on a HATS, acoustic paths are outside the scope of this thesis, although they are currently under study by the author for the development of P.AAM. However, simulations of the transmit path from mouth to network junction are included in most subjective tests and are relevant to this thesis, and three of the tests described in Appendix D include acoustic transmission from HATS to terminal mouthpiece.

Transducers. With a real system the microphone and earpiece of the telephone may have a highly non-ideal frequency response. This is outside the present scope; however, full equalisation in these cases is unlikely to be correct, and this provides motivation for the partial equalisation approach that is adopted in section 4.4.3.

Hybrid and local loop. Most analogue telephones are connected to the exchange or local concentrator by two twisted copper wires. A circuit known as a hybrid allows simultaneous transmission and reception in both directions whilst minimising reflections. The twisted pair acts as a transmission line, with inductive and capacitative effects, and several components in the hybrid act to balance this complex impedance. Due to component variations the frequency response of the combined circuit may not be flat in the passband. For many network testing applications, the test equipment is connected at a 2-wire interface and is also subject to these effects.

Signal conditioning. A current trend in mobile telephones is to perform subband equalisation and noise reduction to maximise speech quality, particularly in noisy environments and with small handsets or headsets that have a microphone well away from the mouth. This may introduce heavy, potentially time-varying, filtering.

4.2.2 Characteristics and models

The send frequency response, which models the transmission from the mouth to the network junction, is the most common linear filter that is included in simulations of networks. Figure 4.1–Figure 4.3 show typical send frequency responses of a number of different telephone devices measured by the author as part of experiment 53 (Appendix D), using an equalised HATS [ITU-T P.58] in an acoustically isolated room and recording the transmitted signal at a digital interface in the network.

Figure 4.1 shows the frequency responses of two PSTN telephone handsets. Figure 4.2 plots the equivalent frequency response for two different headsets connected to a PSTN telephone. In all of these cases the terminal device was located in the standard position, with the mouthpiece in the ear-mouth plane and located just to the side of the mouth. Figure 4.3 shows the frequency response of a market-leading hands-free conference telephone located 0.5m in

front of the HATS. In each case the speech was transmitted wideband through the HATS mouth (using 48kHz sampling rate), and recorded at 8kHz A-law PCM using ISDN. These plots show the frequency response measured using spectral difference, over 96s of speech recordings for each test; cross-spectrum transfer function estimation was found not to be reliable due to sample rate jitter of about 0.01% in the measurements. Analysis was performed at 16kHz sampling rate, using 75% overlapping 512-point Hamming windows for the spectrum estimation.

Figure 4.1: Send response of two handsets

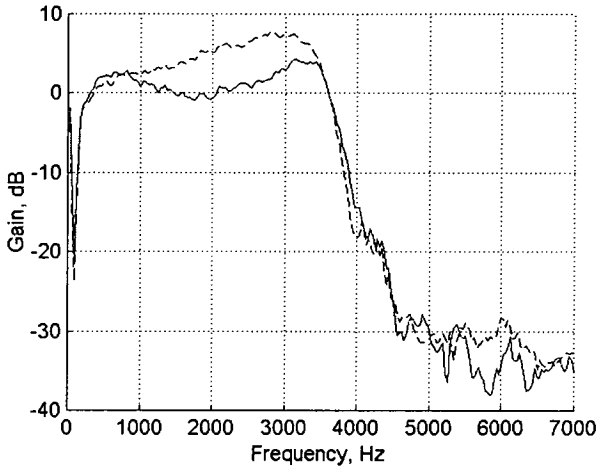


Figure 4.2: Send response of two headsets

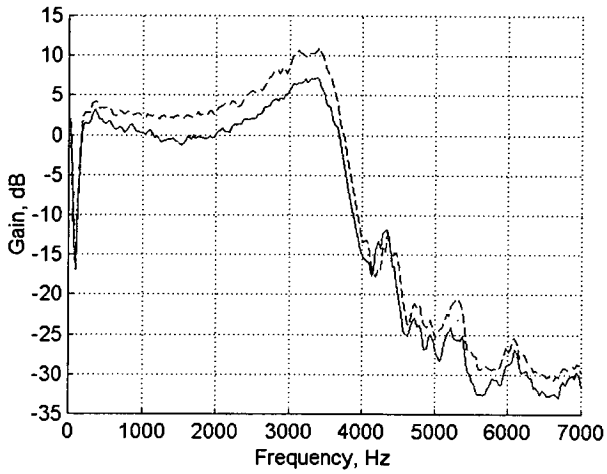


Figure 4.3: Handsfree phone send response

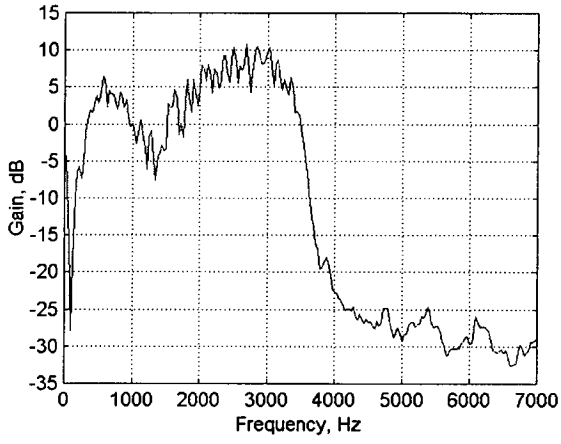
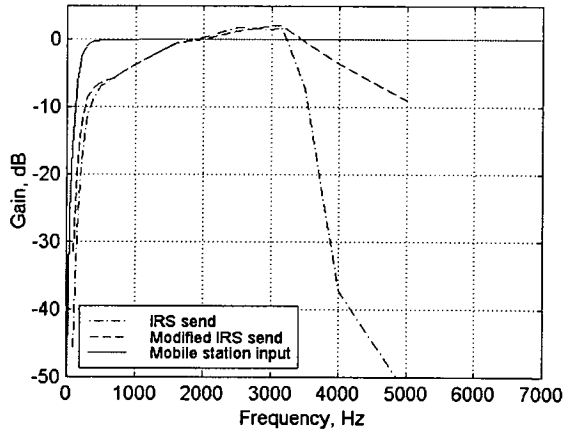


Figure 4.4: Reference handset send models



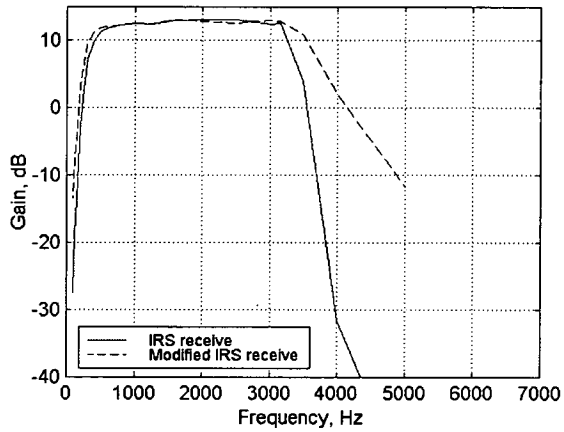
For repeatable subjective testing and network measurement it is much more convenient to use a digital filter with a defined frequency response than to make HATS measurements of physical devices. The ITU-T has therefore standardised a model of the electrical and acoustic components, known as the intermediate reference system (IRS) [ITU-T P.48], with a defined frequency response magnitude. The IRS includes a strongly bandlimiting filter modelling the transmission line, which is not representative of modern networks. A modified IRS characteristic was also produced by removing this filter, and this MIRS send frequency response is commonly

used for processing material for subjective tests [ITU-T P.830]. The IRS and MIRS send frequency responses are shown in Figure 4.4. The filters are usually implemented digitally by zero-phase FIR filters.

The acoustic specifications for mobile telephones are not controlled by the ITU, and the author has found even greater variation in the frequency responses of mobile handsets and headsets. ETSI and the ITU have used a simpler mobile station input filter (MSIN) for subjective tests for speech coders for mobile, with the frequency response of a 2nd-order Butterworth high-pass filter with a –3dB point of 200Hz. This cut-off frequency is lower than the MIRS filter, and the MSIN filter lacks the +10dB/decade boost in the passband that is provided by the IRS and MIRS send filters. The MSIN frequency response is also shown in Figure 4.4.

It is important to note that telephone receivers also show a strong band-pass characteristic. This means that variations in the frequency response outside the 300–3,400Hz passband will in general be inaudible. The standard IRS and MIRS receive filters [ITU-T P.48, ITU-T P.830] are shown in Figure 4.5. Most perceptual models, including PSQM, PAMS and PESQ, include an input filter with a frequency response similar to these [ITU-T P.861, ITU-T P.862].

Figure 4.5: Reference handset receive models



4.2.3 Subjectivity of linear filtering

Subjective test results appear to indicate little difference between the quality of speech filtered through the MIRS and MSIN filters, when used without speech coders. For example, Table 4.1 shows the subjective quality of these two filters, in tandem with G.711 A-law PCM, in a subjective test conducted by the author as part of the P.AAM development. There is no significant difference between the two conditions at the 95% level.

Table 4.1: Subjective effect of filtering

Condition	MOS	95% confidence interval
MSIN–G.711	4.08	±0.16
MIRS–G.711	4.02	±0.15

However, in conjunction with speech coders, the MIRS filter often results in better performance for a given coder and bit-rate. This is thought to be because the greater attenuation of low frequencies of the MIRS filter means that there is more energy at the perceptually important frequencies of 1–3kHz, allowing these to be encoded more accurately. This is illustrated by Table 4.2, which gives results from another subjective test conducted by the author. These conditions compare the GSM-FR coder with the two input filters, in identical radio conditions, using simulated error patterns representative of the given channel SNR (carrier/interference ratio). For 7dB and 13dB channel SNR the quality is significantly higher with the MIRS filter. This was verified for the 13dB case (where the MOS difference is 0.23, and the 95% confidence intervals overlap slightly) by a paired two-sided t-test [Duckworth 1968], for which the probability of the null hypothesis $P(T<t)=0.009$, indicating that it is likely that the means are different.

Table 4.2: Filtering in tandem with speech coder

Condition	MOS	95% confidence interval
MSIN–GSM-FR 13dB channel SNR	3.17	±0.16
MSIN–GSM-FR 7dB channel SNR	2.72	±0.12
MSIN–GSM-FR 4dB channel SNR	1.52	±0.10
MIRS–GSM-FR 13dB channel SNR	3.40	±0.14
MIRS–GSM-FR 7dB channel SNR	2.97	±0.16
MIRS–GSM-FR 4dB channel SNR	1.59	±0.11

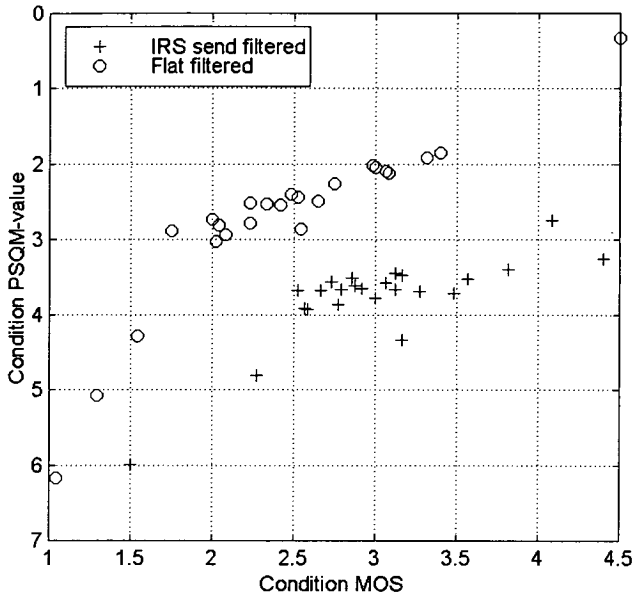
Further examples of the MIRS and MSIN filters, and a number of other filter types (both simulated and measured), are included in the subjective test database used for this thesis. This means that it is possible to evaluate the performance of the perceptual models for a wide range of linear filters in conjunction with coding errors and other distortions.

4.2.4 Performance of PSQM

The effect of filtering on PSQM, which does not include any process for transfer function equalisation, is illustrated by Figure 4.6. This shows a scatter plot of conditions for subjective test 16 (Appendix D), in which half of the conditions were processed through the IRS send filter, and half were unfiltered. The test used the MNRU, GSM-FR and GSM-HR coders. The IRS send filtered conditions are clearly separated from the unfiltered conditions, and the perceptual

model has little predictive power for the main group of filtered coding distortions. The correlation coefficient for this test, after 3rd-order monotonic polynomial regression, is 0.62.

Figure 4.6: Effect of filtering on PSQM



4.2.5 Assumptions

For the remainder of this chapter, a number of assumptions will be made about linear filtering in the system under test.

4.2.5.1 Listening equipment

The recorded signals that are available to the model would be presented to a customer using a narrowband telephone handset with a frequency response similar to the IRS receive characteristic (Figure 4.5). An input filter with this frequency response is therefore used at the start of the perceptual model.

4.2.5.2 Dispersion

As discussed in the previous chapter, the dispersion is assumed to be low, and much smaller than the temporal resolution of the perceptual models. Equivalently, the impulse response is assumed to be short in comparison to the frame size of a perceptual model. In practice this assumption works well for electrical measurements and also for acoustic measurements of handsets on HATS, where the direct path is much louder than any echoes. However this does not hold in the presence of strong echoes, for example with acoustic measurements of hands-free terminals.

This assumption means that the frequency response may be equalised using magnitude-based methods only; full deconvolution is not necessary.

4.2.5.3 Stable

The linear system is assumed to be stable. In other words, its output is bounded for all bounded input values. Unstable systems are generally not suitable for audio transmission.

4.2.5.4 Time invariant

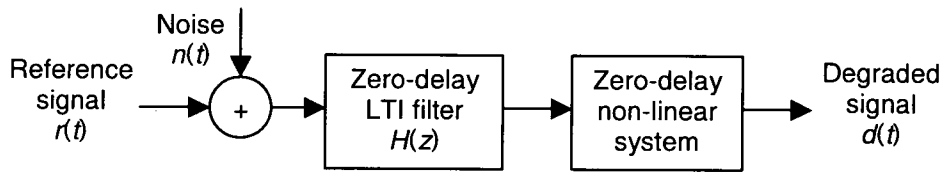
Most perceptual models assume that the frequency response is constant over the duration of the measurement (typically 8s), and average the estimated transfer function over this period. This assumption is valid for most electrical and acoustic interfaces, as these are held constant. However some recent mobile telephones are known to include dynamic frequency response equalisation using three sub-bands. Particularly if there is no run-in period prior to making the test, it is possible that these devices could adapt during a measurement.

A weaker assumption was made by the author for PAMS. Here it was assumed that the frequency response was constant during a speech utterance, and transfer function estimation and equalisation were performed independently for each utterance.

4.2.5.5 Reference model

It is assumed that time-delay estimation has been performed using the methods described in the previous chapter, and that any time-delay has been eliminated before the transfer function or frequency response is estimated. The system can therefore be decomposed into the processes shown in Figure 4.7. For the purpose of this chapter, the LTI filter is to be identified and the frequency response used to equalise the auditory transforms in the perceptual model.

Figure 4.7: System decomposition for transfer function estimation



4.3 Linear transfer function estimation

This section provides an overview of methods for estimating a linear transfer function, where only the input $r(t)$, and output $d(t)$ are known, and for equalising $r(t)$ to $d(t)$. The linear system itself and the corrupting noise $n(t)$ on the observations of $d(t)$ are unknown. Some of the problems of these methods for speech quality assessment are discussed.

4.3.1 Parametric methods for system identification

A large part of the literature in this area focuses on estimation of some vector of parameters of a linear system $H(z)$ of known order and structure, and evaluating whether the postulated order is too low or too high. These are known as parametric methods [Söderström 1989]. For example, the system may be modelled as a K order FIR filter with the addition of uncorrelated random noise $n(t)$ at the output, as shown in (4-1).

$$d(t) = h_0 r(t) + h_1 r(t-1) + \cdots h_{N_k-1} r(t-(N_k-1)) + n(t) \tag{4-1}$$

The method of least squares [Bronshtein 1985] may be applied to minimise the MSE $\frac{1}{N_n}\sum_{t=0}^{N_n-1}(n(t))^2$. By setting the derivative with respect to the filter coefficients to zero, the familiar solution (4-2) may be obtained:

$$\mathbf{h} = (\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T\mathbf{d} \tag{4-2}$$

where

$$\mathbf{h} = \begin{bmatrix} h_0 \\ \vdots \\ h_{N_k-1} \end{bmatrix} \quad \mathbf{d} = \begin{bmatrix} d(0) \\ \vdots \\ d(N_n-1) \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} r(0) & \cdots & r(-N_k+1) \\ \vdots & & \vdots \\ r(N_n-1) & \cdots & r(N_n-N_k) \end{bmatrix} \tag{4-3}$$

This method provides an asymptotically unbiased estimate of $H(z)$ provided that certain conditions are met, in particular that **R** is full rank (which typically requires that $r(t)$ contains energy at all frequencies) and that $N_n >> N_k$. In practice, particularly if the level of the noise $n(t)$ is greater than that of $r(t)$ at some frequencies, or if the order N_k is incorrect, estimation errors on **h** may be large and are not easy to determine. [de Vries 1994] analyses this problem further and suggests techniques to identify the confidence intervals.

4.3.2 Nonparametric methods for transfer function estimation

Methods that estimate properties of the linear system $H(z)$ directly, without relying on some hypothesised model, are known as nonparametric techniques.

4.3.2.1 Spectral difference

Several procedures are available to estimate the spectra $P_{rr}(\Omega)$ and $P_{dd}(\Omega)$ of the signals $r(t)$ and $d(t)$, including the use of filter banks, autocorrelation or spectrum estimation techniques such as Welch's averaged periodogram [Welch 1967, Oppenheim 1989]. Under the assumption that the noise is uncorrelated with the reference signal, the spectra are related by (4-4).

$$P_{dd}(\Omega) = \left|H(e^{j\Omega})\right|^2 P_{rr}(\Omega) + P_{nn}(\Omega) \tag{4-4}$$

This relationship is often applied, ignoring the noise, to derive a simple estimate of the magnitude of the frequency response $\left|\hat{H}(e^{j\Omega})\right|^2$ (4-5). This method was used for the results shown in Figure 4.1–Figure 4.3.

$$\left|\hat{H}(e^{j\Omega})\right|^2 = \frac{P_{dd}(\Omega)}{P_{rr}(\Omega)} \tag{4-5}$$

In practice, the estimate using (4-5) is biased if noise is present, and cannot fall below the noise-to-signal ratio at any given frequency: its relationship with the actual transfer function is shown in (4-6). If the noise spectrum is unknown – as is the case for this thesis – it is impossible to eliminate the error term $P_{nn}(\Omega)/P_{rr}(\Omega)$. Despite this bias, spectral difference is popular in many applications because of its simplicity and robustness to non-linear or time-varying processes such as clock jitter.

$$\left|\hat{H}(e^{j\Omega})\right|^2 = \left|H(e^{j\Omega})\right|^2 + \frac{P_{nn}(\Omega)}{P_{rr}(\Omega)} \tag{4-6}$$

4.3.2.2 Transfer function estimation using the cross-spectrum

Under the assumption that the system under test is stable and LTI, an estimate of $H(z)$ may be obtained with much better noise rejection by using the cross-spectrum. Most textbooks perform the following derivation using time-domain cross-correlation, which is more involved but does not require the assumption of wide-sense stationarity [Söderström 1989; Ljung 1987].

The cross-spectrum between $r(t)$ and $d(t)$ may be written as follows.

$$P_{rd}(\Omega) = \mathrm{E}[R^*(e^{j\Omega})D(e^{j\Omega})] \tag{4-7}$$

In (4-7), * denotes the complex conjugate and E[] is the expectation operator. The system is represented by (4-8).

$$D(z) = H(z)R(z) + N(z) \tag{4-8}$$

Assuming that the signals are wide-sense stationary – effectively that the expectation is taken over many finite duration instances of the processes – and substituting into (4-7):

$$P_{rd}(\Omega) = \mathrm{E}[R^*(e^{j\Omega})\{H(e^{j\Omega})R(e^{j\Omega}) + N(e^{j\Omega})\}] \tag{4-9}$$

$$P_{rd}(\Omega) = H(e^{j\Omega})\mathrm{E}[R^*(e^{j\Omega})R(e^{j\Omega})] + \mathrm{E}[R^*(e^{j\Omega})N(e^{j\Omega})] \tag{4-10}$$

Under the further assumption that the noise is independent of and uncorrelated with $r(t)$, the second term of (4-10), $P_{rn}(\Omega)$, is zero, giving

$$P_{rd}(\Omega) \approx H(e^{j\Omega})P_{rr}(\Omega) \tag{4-11}$$

From (4-11), an estimate of the complex transfer function using the ratio of the cross-spectrum to the spectrum of the input signal can be derived (4-12). These spectra can be efficiently computed using the averaged periodogram [Welch 1967, Oppenheim 1989]. The inverse DFT of $\hat{H}(e^{j\Omega})$ gives an estimate of the impulse response $h(t)$.

$$\hat{H}(e^{j\Omega}) = \frac{P_{rd}(\Omega)}{P_{rr}(\Omega)} \tag{4-12}$$

It is worth noting the conditions under which (4-12) provides a good estimate of the transfer function when used in conjunction with the periodogram [Söderström 1989; de Vries 1994]:

- The system must be strictly time-invariant
- The impulse response $h(t)$ must lie well within the frames used for the cross-spectrum estimation
- The noise must be uncorrelated with the input signal
- The estimation variance of $\hat{H}(e^{j\Omega})$ is inversely proportional to the number of periodograms averaged
- If the SNR is negative at some frequencies, the noise cross-spectrum component $P_{rn}(e^{j\Omega})$ is large, leading to a large variance on $\hat{H}(e^{j\Omega})$ that may not be sufficiently reduced by periodogram averaging.

4.3.3 Coherence function

A related quantity to the transfer function estimate is the magnitude squared coherence of the two signals $r(t)$ and $d(t)$, defined as follows [Knapp 1976].

$$C_{rd}(\Omega) = \frac{|P_{rd}(\Omega)|^2}{P_{rr}(\Omega)P_{dd}(\Omega)} \tag{4-13}$$

In the noiseless case $C_{rd}(\Omega)=1$. As noise or non-linear effects rise, the coherence falls. However, low coherence does not imply that $\hat{H}(e^{j\Omega})$ is necessarily wrong, because the averaging process, particularly over large numbers of frames, provides good rejection of noise even though the coherence may still be low.

4.3.4 Effect of non-linearity and time variance

Several properties of the communications systems that are of interest for this thesis have a significant impact on the estimates described in this section. The problems are essentially the same as those discussed in Chapter 3.

Clock jitter can be a problem where two separate devices are used to inject the reference signal $r(t)$ and capture the degraded signal $d(t)$, as it is prohibitively expensive to precisely synchronise the sample rate clocks used to determine the playout and recording rates. A typical clock stability is 0.01%, which means that over a 10s measurement an offset of up to 1ms may accumulate. While the time-delay estimation techniques described in the previous section may reduce this, they cannot eliminate it because of the assumption of piecewise constant delay.

Resynthesising coders and other non-linear elements may destroy the short-term linear relationship between $r(t)$ and $d(t)$. For example, an unvoiced sound may be re-synthesised in the decoder by shaping locally-generated noise with the spectral envelope of the reference. In this case the two signals are incoherent. Figure 3.2 presented an example of the effect of these phenomena on linear transfer function estimates.

These processes are most critical for the estimation of the cross-spectrum $P_{rd}(\Omega)$, as illustrated in the next section. The clock stability problem may be largely solved by using a high-resolution frequency-domain delay estimate [Johnson 1993] to follow the clock; however, this is beyond the scope of this thesis.

4.3.5 Example results

A simulated telephone connection, shown in Figure 4.8, provides an illustration of the strengths and weaknesses of these techniques. The reference signal was an 8s sentence pair spoken by a male talker. Noise recorded in an office environment was added, at various SNRs, at the start of the processing, the standard method for testing the performance of codecs and networks in noise. The linear filter used was an IIR implementation of the MIRS send characteristic [ITU-T P.830] developed by the author using a log-magnitude method similar to that of [Lin 2001], with the impulse response shown in Figure 4.9. The signal was down-sampled to 8kHz and passed through a number of speech codecs, including channel error models. As part of the up-sampling process, spline interpolation was used to allow clock jitter to be modelled. Note that because the system operates at 8kHz sampling rate, the impulse response in Figure 4.9 includes bandlimiting, and in the following figures all of the estimation methods can be expected to identify strong attenuation above 4kHz.

Figure 4.8: Simulation framework for transfer function estimation

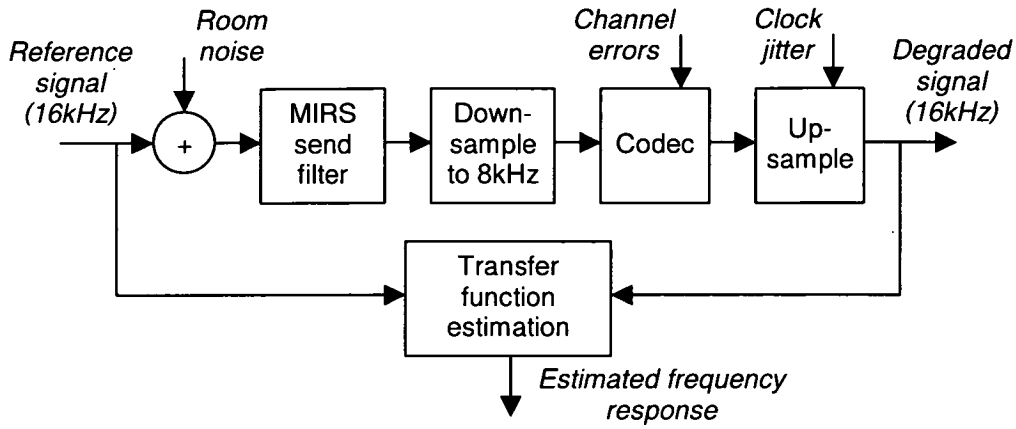
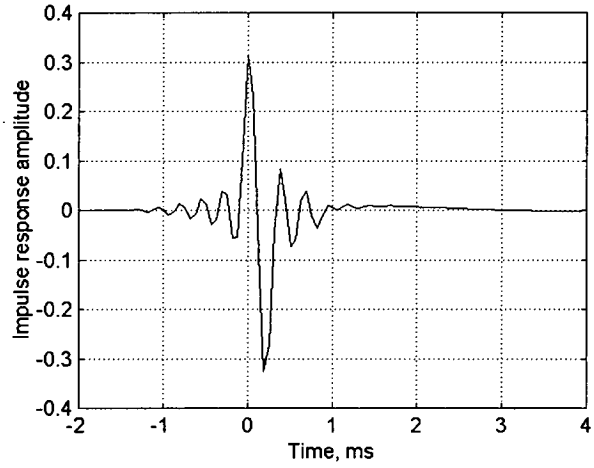


Figure 4.9: MIRS send filter impulse response



4.3.5.1 Linear system

With no clock jitter and with G.711 A-law as the codec, the system is effectively linear and is time-invariant (G.711 performs non-linear quantisation, but the transfer function estimation effectively treats the quantisation error as noise).

A parametric estimate of the system was computed using minimum MSE (4-2) to identify a 49-tap FIR filter centred on zero delay. The time-domain filter estimate is shown against the "true" impulse response from Figure 4.9 – including bandlimiting due to sample rate conversion – in Figure 4.10. The corresponding frequency response is shown in Figure 4.11. These were computed at 0dB SNR; the parametric method provides excellent noise rejection and in fact the results are indistinguishable from those without noise. For this filter order, the parametric method is underdetermined but does provide a fairly good estimate, with about 3dB of ripple within the passband.

Figure 4.10: Parametric impulse response estimate

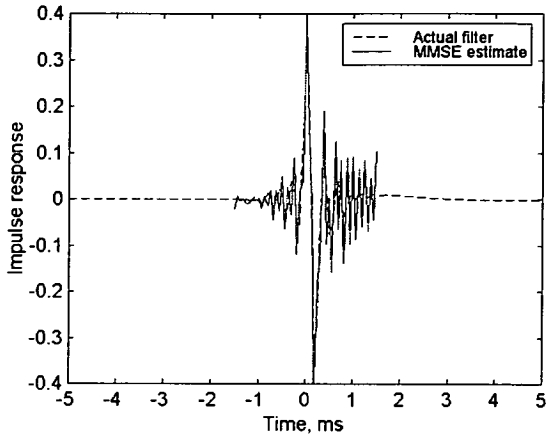
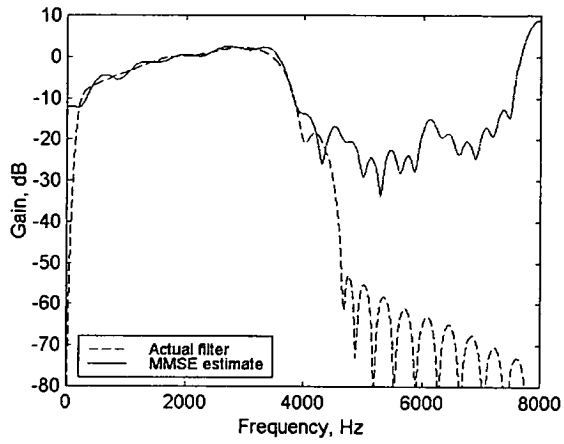


Figure 4.11: Parametric frequency response estimate



Nonparametric estimates of the transfer function were computed using Welch's averaged periodogram for spectrum estimation, with 50% overlapping Hann windows of 512 points (32ms). The frequency response estimates using spectral difference (4-5) and cross-spectrum transfer function estimation (4-12) are shown in Figure 4.12 and Figure 4.13, also for the 0dB SNR case. The spectral difference estimate is quite seriously affected by the noise, particularly below 2kHz where the noise spectrum starts to exceed the speech spectrum; the estimate becomes closer if the noise level is reduced. For the cross-spectrum transfer function estimate, the noise causes a smaller amount of ripple – about 1dB in the passband rising to 10dB at 200Hz. At better SNR the cross-spectrum method provides an even more accurate estimate of the transfer function. The cross-spectrum transfer function estimate is much more accurate in this example than the parametric estimate largely because the latter has too few degrees of freedom for this case.

Figure 4.12: Spectral difference frequency response estimate, 0dB SNR

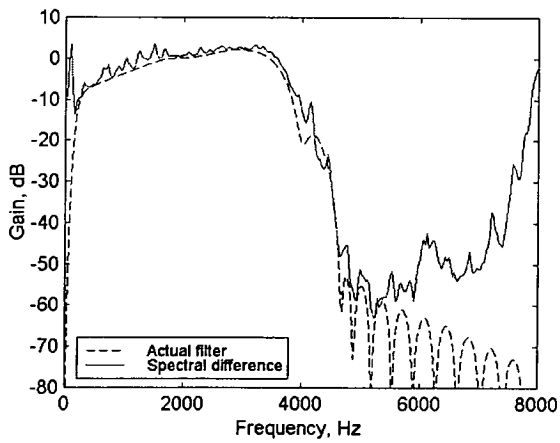
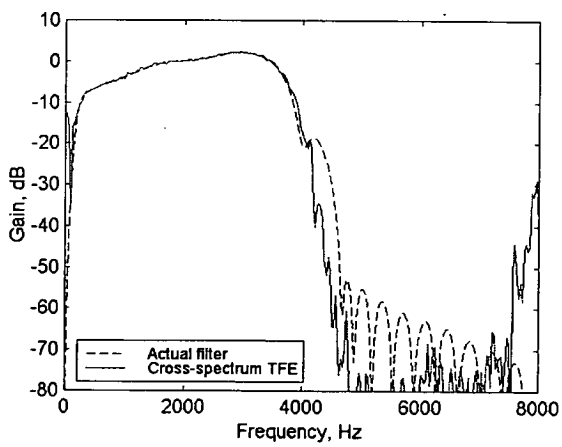


Figure 4.13: Cross-spectrum transfer function estimate, 0dB SNR



4.3.5.2 Effect of clock jitter

The effect of clock asynchrony in the transmit and receive device was simulated by interpolating the degraded signal, using spline interpolation, prior to up-sampling. The following examples show the effect of the receive clock running 0.01% too fast – gaining 0.8ms over the 8s duration of this test – at 20dB SNR. The codec is G.711 A-law.

The parametric and cross-spectrum frequency response estimates, shown in Figure 4.14 and Figure 4.15 respectively, are both severely affected by the clock jitter. The spectral difference estimate, shown in Figure 4.16, is largely unchanged by this small amount of jitter; the main error compared to the jitter-free case is a drop of about 1dB between 2.5–3.5kHz. The coherence function, shown in Figure 4.17, shows that the jitter causes a large drop in correlation between the reference and degraded signals when averaged over the duration of the measurement, and this is the cause of the large estimation errors with the parametric and cross-spectrum methods, which both rely on the linearity of the system.

Figure 4.14: Parametric frequency response estimate, 0.01% clock jitter

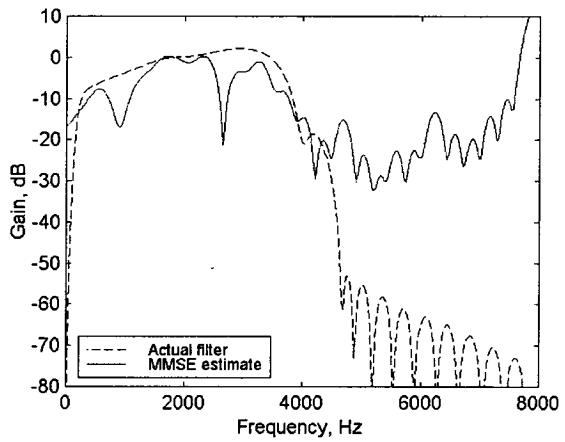


Figure 4.15: Cross-spectrum TFE, 0.01% clock jitter

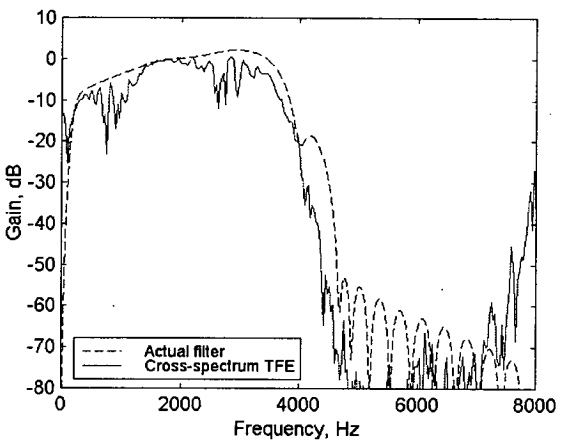


Figure 4.16: Spectral difference frequency response estimate, 0.01% clock jitter

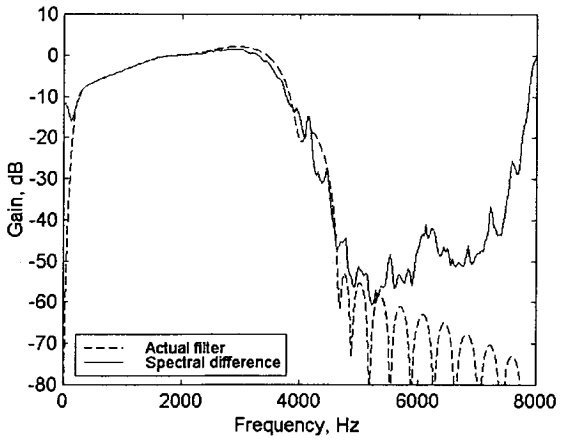
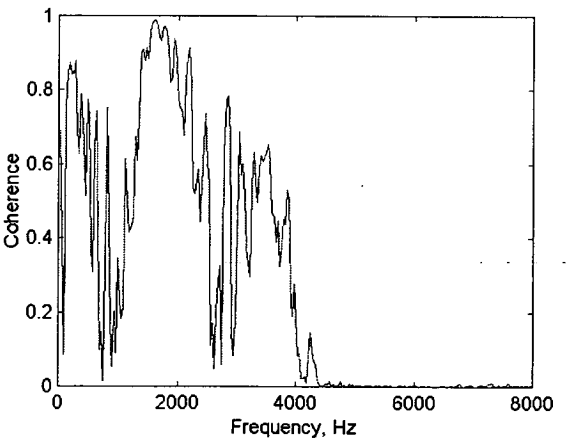


Figure 4.17: Effect of clock jitter on coherence



4.3.5.3 Effect of coding distortions

Low bit-rate coders do not necessarily preserve the detailed temporal structure of the waveform, particularly in the presence of channel errors. The effect of this is shown by the following examples, which are based on GSM-EFR (12.2kbit/s) [GSM 06.60] and G.723.1 at 5.3kbit/s [ITU-T G.723.1], both at 20dB SNR.

Figure 4.18 and Figure 4.19 plot the magnitude of the cross-spectrum transfer function estimate for EFR in 16dB channel SNR (good conditions) and 7dB channel SNR (poor conditions). The corresponding TFE for G.723.1 in 3% frame erasure is shown in Figure 4.20. The effect of coding distortions and errors is to reduce the coherence, making the cross-spectrum underestimate the "true" frequency response. The coding distortion causes the spectral difference, shown in Figure 4.21 for G.723.1 in 3% frame erasure, to be noisier than for G.711, but it is more accurate in the passband than the linear estimates, although it does show low-frequency bias due to noise in a similar way to Figure 4.16.

Figure 4.18: Cross-spectrum TFE, EFR, 16dB channel SNR

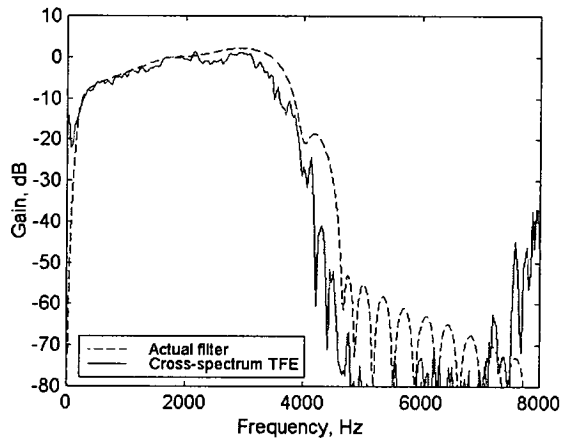


Figure 4.19: Cross-spectrum TFE, EFR, 7dB channel SNR

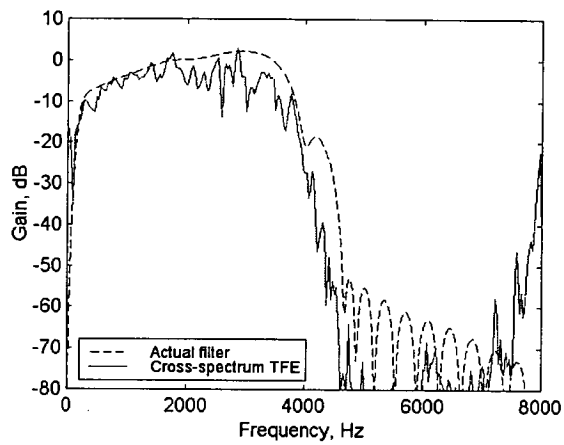


Figure 4.20: Cross-spectrum TFE, G.723.1 5.3kbit/s, 3% frame erasure

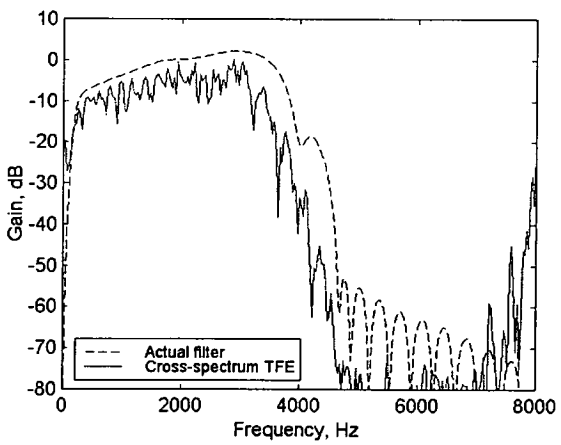
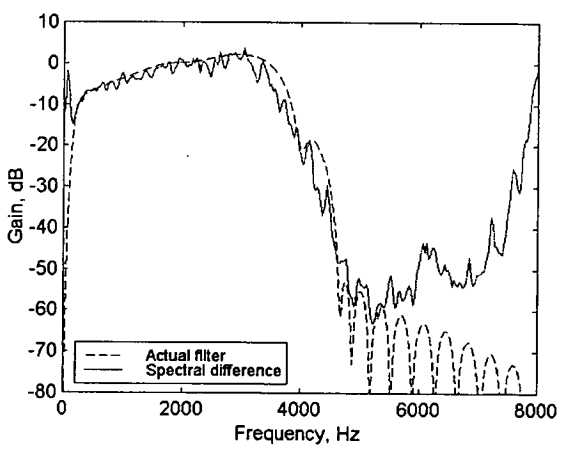


Figure 4.21: Spectral difference, G.723.1 5.3kbit/s, 3% frame erasure



4.4 Perceptual transfer function equalisation

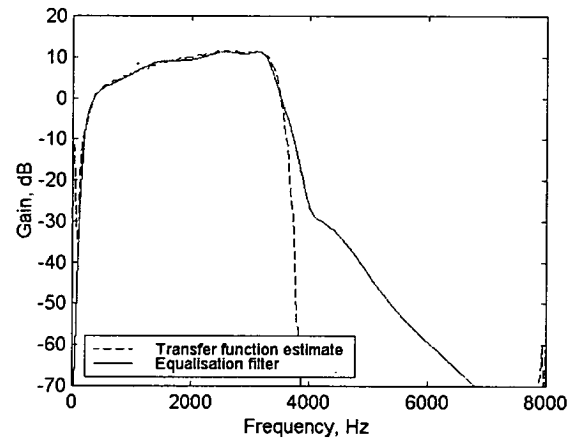
The previous section has shown that transfer function estimation methods may be used to find an approximation to the frequency response of the linear part of the system under test. This section considers the application and generalisation of these techniques to improve the accuracy of perceptual models in the presence of linear filtering.

4.4.1 Objective

The purpose of this chapter is to improve the accuracy of perceptual models with conditions that include linear filtering. Following [Thiede 1996], this is achieved by equalising the reference signal to the degraded signal.

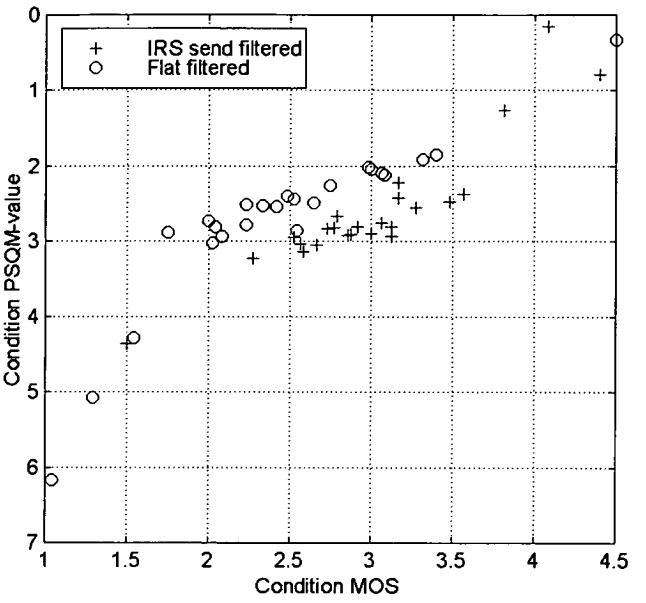
Figure 4.6 showed that PSQM, with no frequency response equalisation process, gives highly inaccurate predictions for conditions that include filtering. The data presented in that example was based on simulations, and includes an unprocessed reference condition that allows the send filter to be checked. The filter was found to be very similar to the standard IRS send characteristic, shown in Figure 4.4 [ITU-T P.48]. An IIR implementation of this filter was designed by the author; the filter impulse response was presented in Figure 3.6. The frequency response estimated using the cross-power method for one of the high-quality conditions, and the response of this filter, are plotted in Figure 4.22.

Figure 4.22: Input filter estimate and equalisation filter



This filter implementation was applied to the reference signal for all of the conditions in the test where the degraded signal had been IRS send filtered, and the data was then processed again through PSQM. The results are presented in Figure 4.23. The correlation coefficient for the equalised material, after 3rd-order monotonic polynomial regression, is 0.87, which compares to 0.62 without equalisation (Figure 4.6). While the equalisation has not eliminated the offset between the filtered and unfiltered conditions, it has reduced it substantially.

Figure 4.23: PSQM with equalisation of reference



Equalisation of the reference signal clearly improves the accuracy of perceptual models in the presence of linear filtering. However, unlike the example shown in Figure 4.23, the filter is generally unknown and must be estimated, despite the presence of noise, clock jitter, low bit-rate coding and channel errors. The proposed modification to perceptual models therefore consists of the following stages.

- Perceptual transfer function estimation
- (Partial) equalisation of the reference to the degraded signal.

The concept of partial equalisation is important. Strong filtering, such as spikes in the frequency response or bandlimiting with cut-off frequencies between about 500 and 3000Hz [Voran 1997], produces a significant drop in perceived quality. Full equalisation in this case is wrong, as it would cause the model to give quality scores that are too high.

This can be addressed either by computing a separate distortion parameter, or by a partial equalisation process. PEAQ performs full equalisation, but includes a measure of the loudness of the signal that has been lost due to filtering as a distortion parameter used in the prediction of quality [BS.1387]. For speech quality measurement, for the reasons introduced in the previous section, the author found that it is difficult to perform accurate full equalisation, and partial equalisation using perceptual bands, described in section 4.4.3, provides a way to include the effect while allowing more robust transfer function estimation [Rix 1999f, Beerends 2002]. This is an application of the partial equalisation concept that was used for time-domain local scaling in PSQM, which was described in section 2.5.2.

In a similar way to [Thiede 1996], it was found most convenient to implement transfer function estimation and equalisation as part of the auditory transform. At this point in the models, the time-delay is known and a perceptual time-frequency transform is performed, which can be readily extended for this purpose.

4.4.2 Other authors' approaches

4.4.2.1 Thiede

The only perceptual model in the literature prior to this study that included a transfer function equalisation process was DIX [Thiede 1996], which is designed for the assessment of audio coders. The auditory transform and equalisation process of this model were subsequently incorporated with few changes into PEAQ [ITU-R BS.1387]. As discussed in the introduction to this chapter, the equalisation process in DIX acts to eliminate three main effects: bandlimiting, which is highly audible in audio quality assessment; small amounts of filtering due to analogue interconnections; and slow variations in gain.

The auditory transform in DIX uses a complex filterbank similar to that described in section 2.5.5. This performs time-frequency analysis and models masking in time and frequency through spreading and smoothing processes. The smoothed power in each perceptual band, termed the excitation, is used for the equalisation process.

The excitation is further averaged in time and frequency. Each band is smoothed over time by a first-order filter with time constant varying from 100ms at low frequencies to 8ms at high frequencies. The ratio between these smoothed spectra provides a correction factor for each signal: if the reference is louder than the degraded signal, it is attenuated and the degraded signal is unchanged; if the reference is quieter, then the degraded signal is attenuated and the reference is unchanged. These correction factors are further smoothed over time using the same first-order filter, and then over frequency by averaging over rectangular windows of about 1 bark width, before they are applied to equalise the signals to each other.

In DIX, an estimate of the spread of the correction factors was used as a distortion parameter; however, this was not included in the prediction of audio quality. In PEAQ, the signal bandwidths are used in the basic version, and an average of the loudness of the signal components that have been lost through equalisation is used in the advanced version.

4.4.2.2 Berger

Only a very brief overview of Berger's model, TOSQA, has been published [Berger 1997]. According to this, the estimated frequency response of the system under test is used to equalise the reference signal, after the STFT, to the degraded signal. No separate distortion

parameter is used to model the effect of the filtering. No information is available on how the frequency response is computed or whether partial equalisation is performed.

4.4.2.3 Park

Based on the ideas presented by the author on transfer function equalisation for PAMS [Rix 1999f], Park developed a modification to BSD [Park 2000]. This applies the coherence function (4-13) to the sone loudness in each bark band, where $L_{rd}(b)$ is the phaseless cross-spectrum between the two signals in the bark, sone domain for band b, and $L_{rr}(b)$ and $L_{dd}(b)$ are the corresponding bark spectra, giving a bark coherence function *BCF* (4-14). The overall quality score, bark distortion-to-signal ratio (BDSR), is computed using (4-15).

$$BCF(b) = \frac{L_{rd}(b)^2}{L_{rr}(b)L_{dd}(b)} \tag{4-14}$$

$$BDSR = \sum_b \left[\frac{1}{BCF(b)} - 1 \right] \tag{4-15}$$

BSD uses deeply-overlapping perceptual bands. This approach therefore effectively performs partial equalisation in a similar way to the method that the author applied to PAMS.

While (4-14) and (4-15) are attractively simple, they do not take account of the relative loudness of each perceptual band, and will show a significant drop in quality in the presence of time-varying gain. In addition, the asymmetry effect is not modelled, so the model may be less accurate for conditions including additive noise.

4.4.2.4 Beerends and Hekstra

For PESQ, Beerends and Hekstra implemented a bark spectrum equalisation process similar to that presented by the author [Rix 1999f], but using spectral difference only [Beerends 2002]. The spectra over the whole signals are calculated using the rectangular critical bands of the auditory transform, with minimal overlap. Because of the band shape used, there is no smoothing between bands. Partial equalisation is implemented by constraining the transfer function estimate to ±20dB and rolling it to unity as the signals fall below a threshold, and this is used to equalise the reference signal to the degraded signal. Local scaling to equalise time-varying gain is performed after the transfer function equalisation has been applied; this acts to remove any constant offset in the transfer function due to estimation errors.

4.4.3 Perceptual frequency response equalisation

The examples presented in Figure 4.11–Figure 4.21 show that standard techniques for transfer function estimation may not perform well in the presence of low bit-rate coding and noise. The

nonparametric methods must use quite long frames in order to provide sufficient resolution at low frequencies. Their resolution is generally equal at all frequencies. This means that they have many degrees of freedom at higher frequencies and can all produce highly rough estimates in this region. It is implausible that the ear could detect and equalise filtering with equal frequency resolution: there is considerable evidence that the bandwidth of channels in the human auditory system increases in proportion to frequency – decreasing the frequency resolution. This has been modelled by the bark frequency scale and variants such as equivalent rectangular bandwidth [Moore 1997a]. These scales are used for perceptual quality assessment, and it seems compelling to use the same perceptual basis for transfer function estimation as part of a perceptual model.

Two nonparametric methods for perceptual transfer function estimation will be explored here, based on spectral difference and the cross-power spectrum. Both of these methods yield an estimate of the magnitude of the frequency response, but no phase information, and can therefore not be used for impulse response estimation. An alternative method that provides an approximation to the perceptual frequency scale but can be used with parametric system identification methods to estimate the impulse response is the warped Z transform [Härmä 2000]. Because the perceptual models considered in this thesis do not take account of phase this was not considered further.

It is assumed for this section that a perceptual representation of signal power in frequency bands is available.

4.4.3.1 Spectral difference

It was found in section 4.3 that spectral difference provides a simple estimate of the frequency response and is fairly robust to typical coding distortions, although it is affected by noise. Estimation of the spectral difference using perceptual methods is straightforward: the power spectra of $r(t)$ and $d(t)$ are averaged over the signal duration either using a perceptual filterbank or by transformation from the STFT, and the ratio gives an estimate of the magnitude squared frequency response according to (4-5), evaluated in the perceptual bands.

This process is equivalent to:

- convolution of the power spectrum of reference and degraded signals with a frequency-variant spreading function
- division in the frequency domain according to equation (4-5).

4.4.3.2 Phaseless cross-spectrum

The perceptual models under consideration for this thesis all produce a representation of signal power in time and frequency, either by a filterbank or using the STFT. (Note that this is not the

case for models based on lower-level neural models such as Meddis's hair cell model [Hauenstein 1998].) From this, a modified cross-spectrum may be computed which discards the phase component. Although this does not preserve the time-domain cross-correlation relationship that is encoded by the standard cross-spectrum of (4-7), it is nevertheless useful for transfer function estimation. In the following, the notation $D(e^{j\Omega})$ is taken to be the STFT of a windowed section of the signal, or alternatively the RMS signal level in the given perceptual band. The phaseless cross-spectrum is defined by (4-16).

$$\Phi_{rd}(\Omega) = \mathrm{E}[\left|R^*(e^{j\Omega})D(e^{j\Omega})\right|] \tag{4-16}$$

Substituting using (4-8) gives (4-17), where the approximation is derived using the triangle inequality, and tends to equality as $\left|N(e^{j\Omega})\right| \to 0$.

$$\begin{aligned}\Phi_{rd}(\Omega) &= \mathrm{E}[\left|R^*(e^{j\Omega})\right|\left|H(e^{j\Omega})R(e^{j\Omega}) + N(e^{j\Omega})\right|] \\ &\leq \left|H(e^{j\Omega})\right|\mathrm{E}[\left|R^*(e^{j\Omega})\right|\left|R(e^{j\Omega})\right|] + \mathrm{E}[\left|R^*(e^{j\Omega})\right|\left|N(e^{j\Omega})\right|]\end{aligned} \tag{4-17}$$

Assuming that $\left|N(e^{j\Omega})\right| = 0$, (4-17) can be re-arranged and the reference signal spectrum substituted to give an estimate of the magnitude of the frequency response (4-18), analogous to (4-12). As argued above, the perceptual models of interest discard phase, so an estimate of the frequency response magnitude is sufficient for the purpose of transfer function equalisation.

$$\left|\hat{H}(e^{j\Omega})\right| = \frac{\Phi_{rd}(\Omega)}{\Phi_{rr}(\Omega)} \tag{4-18}$$

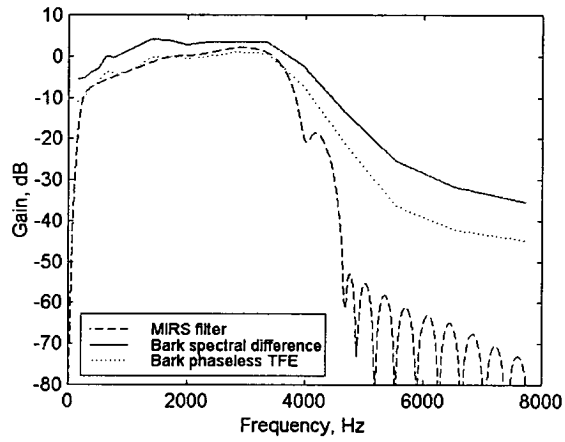
The noise sensitivity may be compared to the spectral difference method (4-6) by substitution of (4-17) into (4-18), to give

$$\left|\hat{H}(e^{j\Omega})\right| \leq \left|H(e^{j\Omega})\right| + \frac{\Phi_{rn}(\Omega)}{\Phi_{rr}(\Omega)} \tag{4-19}$$

where $\Phi_{rn}(\Omega)$ is the phaseless cross-spectrum between reference and noise, which does not in general converge to zero. In the case of uncorrelated Gaussian processes $r(t)$ and $n(t)$, (4-18) provides a 3dB improvement over the spectral difference method (4-6). In practice for speech signals a larger improvement may be achieved because of the amplitude statistics of the short-term speech spectrum. This is illustrated by Figure 4.24, which shows the transfer function estimates using these two methods for the example system introduced in the previous section, for all of the main distortion effects combined. The SNR is –10dB SNR, clock jitter is at 0.01 and the speech is coded with EFR at 7dB channel SNR. The PAMS bark filterbank was used for the time-frequency analysis. Within the passband the bark phaseless TFE is 3–5dB lower than the spectral difference estimate, and is within 2dB of the linear frequency response in

the range 300–3,400Hz. Above 4kHz, the phaseless method provides about 10dB better rejection of noise, and is mainly limited by the shape of the perceptual filters, which limit the slope in the frequency domain.

Figure 4.24: Perceptual transfer function estimates



4.4.3.3 Higher-order phaseless cross-spectrum

The approach described above may be generalised to higher orders by modifying (4-16) to include optional powers for each signal in the phaseless cross-spectrum (4-20). For certain noise and signal statistics this may provide an improvement. For example, with $a>1$, more weight is given to louder parts of the reference, decreasing the noise sensitivity.

$$\Phi_{rd}^{a,b}(\Omega) = \mathrm{E}[\left|R^*(e^{j\Omega})\right|^a \left|D(e^{j\Omega})\right|^b]^{\frac{1}{ab}} \tag{4-20}$$

In practice minimal improvement was found using this generalisation because of the diverse range of noise and errors that may be encountered. Sensitivity to loud distortions that coincide with loud parts of the reference, such as impulsive noise, increases if $b>1$. $a>1$ weights the computation towards the loudest sections of the reference; although this reduces the noise bias, it effectively uses a smaller proportion of the signals in the computation and therefore increases the estimation variance. Sensitivity to stationary noise increases if either $a, b<1$. The absolute magnitude cross-spectrum (4-16) seems to give a good balance between these effects.

4.4.3.4 Perceptual smoothing

Sharp resonances or transitions, or strong bandlimiting, in the frequency response may be audible, and full equalisation in these cases can lead to a perceptual model giving quality scores that are too high [Ordas 2001]. This may be addressed by smoothing the short-term frequency response of the system using perceptual methods, and effectively means that partial equalisation is performed.

With STFT-based models such as PESQ, perceptual smoothing may be implemented using smearing in the frequency domain in a similar way to masking models [Beerends 1992; ITU-R BS.1387]. This may be performed either in linear frequency, using a frequency-dependent perceptual filter shape, or in the bark domain using low-pass filtering with cut-off rates on the order of 10–25dB/bark upward spread of masking, 30dB/bark lower spread [Theide 1996]. The sum of the two spreading rates provides an upper limit on the rate of change of the frequency response estimate.

Some filterbank-based models, including PAMS, DIX and PEAQ, perform smoothing automatically because they are based on deeply-overlapping perceptual filters. However, it was found that the performance of PAMS was improved by performing further smoothing, in a similar way to DIX and PEAQ, using a moving average to smooth the spectral estimates in the frequency domain.

4.4.3.5 Time-varying distortions and coherence weighting

The author found that a weakness of all the methods described in this section was that they may show significant bias under certain classes of distortion. There are two typical problem conditions.

The first is a mobile connection with bit errors that are not corrected or detected by the channel coder. This can cause broadband speech-like sounds to be generated at arbitrary levels; in some types of poor radio condition, loud distortions during speech and silence can be common. Because these sounds are loud, they may significantly bias the degraded signal spectrum, making spectral difference overestimate the frequency response. Loud, speech-like distortions are also more common during speech utterances due to long-term level prediction in speech coders; because of correlation between the distortions and speech, the phaseless cross-spectrum method provides little rejection of these distortions and may also overestimate the frequency response.

The second is muting under frame erasure. VoIP and certain mobile conditions can cause prolonged periods where the degraded signal is muted. The cross-spectrum and the degraded signal spectrum are both reduced by muting, leading to underestimation of the frequency response magnitude.

A pragmatic solution to these two problems developed by the author for PAMS was to weight the spectrum and cross-spectrum estimates by the modified local signal coherence $c_l(k)$, defined by (4-21)

$$c_l(k) = \left(\frac{\sum_f \sqrt{r_s(k,f) d_s(k,f)}}{\sum_f \max(r_s(k,f), d_s(k,f))} \right)^{\lambda} \tag{4-21}$$

where $r_s(k,f)$ and $d_s(k,f)$ are the reference and degraded signal excitation (in units of power), for frame k and perceptual band f, and $\lambda>0$ is an exponent that controls the strength of the weighting. The maximum value is used on the denominator, rather than a conventional coherence measure such as (4-13), to ensure that low weight is given in the event of a large mismatch in levels, for example due to additive noise.

The weighting is used in the averaging over k to calculate the relevant signal or cross-spectrum. Assuming that highly distorted frames will have lower $c_f(k)$ than undistorted frames, the frequency response estimate will therefore be less biased by time-varying degradations.

The reduction in bias by using local coherence weighting with spectral difference estimates in the presence of severe time-varying distortion is illustrated by Figure 4.25 and Figure 4.26. This condition is based on the reference system with a known linear filter introduced in Figure 4.8. In this case, both background noise from a street scene, and impulsive noise (modelling bit errors in G.711 transmission) were introduced as distortions. Figure 4.25(a) plots the function calculated from the bark perceptual transform using (4-21), with $\lambda=1$, which gives most weight to the speech sections. The degraded signal $d(t)$ is shown in Figure 4.25(b). The frequency response estimates calculated using (4-5) are presented in Figure 4.26, showing that the weighting process provides much greater noise rejection.

Figure 4.25: Local coherence

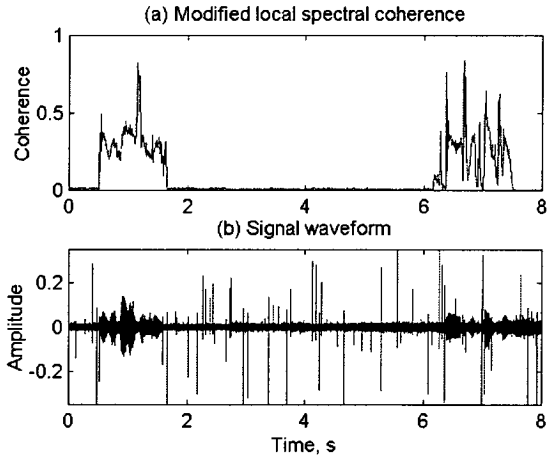
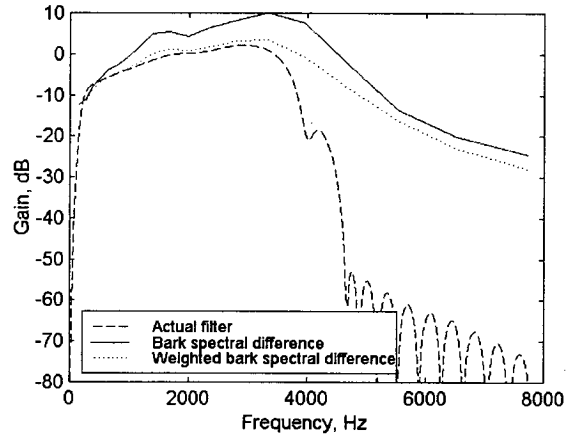


Figure 4.26: Bark spectral difference with local coherence weighting



4.4.3.6 Perceptual frequency response equalisation

Following [Thiede 1996], the reference signal is equalised by the (smoothed) frequency response estimate of the system under test. As was shown by the example of Figure 4.23, this improves perceptual model performance by substantially reducing the amount of distortion measured due to the filtering.

A modification to this method was developed by Beerends and Hekstra for PESQ [Beerends 2002]. The spectral difference method that they implemented operates in rectangular critical bands, without perceptual smoothing. To deal with the problem of sharp changes in the frequency response, they constrain the amount of equalisation that is performed to ±20dB. The frequency response estimation in PESQ is performed from pesq_psychoacoustic_model(), which calls time_avg_audible_of() to compute the spectra of each signal, and freq_resp_compensation() to partially equalise the reference signal to the degraded signal. All of these functions are in file pesqmod.c [ITU-T P.862].

Ordas and Fox have criticised this as being too broad a range, providing a counter-example with a rapid ±15dB swing in the passband at about 1kHz [Ordas 2001] that receives a PESQ score of 4.43, only very slightly below the maximum of 4.5. They assert that this is a highly audible distortion, which should be penalised more heavily, although the correct quality score in this case is a matter for debate.

4.5 Results

The methods described in this chapter were evaluated by modifying the processing in PESQ. In addition, PAMS and PSQM were also processed through the same data. The models compared are as follows.

(A) PSQM [ITU-T P.861] extended with constant-delay histogram time alignment described in section 3.4 (no frequency response compensation).

(B) PESQ with frequency response compensation disabled.

(C) PAMS 3.1 (January 2001), which performs partial transfer function equalisation in the bark filterbank using a combination of spectral difference and phaseless cross-spectrum estimation, with frequency-domain smoothing and local coherence weighting.

(D) PESQ, spectral difference evaluated using (4-5) in the bark critical band (pitch) domain.

(E) PESQ as standardised in [ITU-T P.862].

(F) PESQ with perceptual smearing at –30dB/bark using phaseless cross-spectrum transfer function estimation (no coherence weighting).

(G) As (F), but with coherence weighting using (4-21) with λ=1.

(H) As (F), but with coherence weighting using (4-21) with λ=2.

(I) As (H) but using spectral difference.

These models are compared in Table 4.3 and Table 4.4. Table 4.3 presents the correlation coefficient, calculated as described in section 2.6, for four individual subjective tests. Test 16

contains two different types of filter, and was used for the examples shown in Figure 4.6 and Figure 4.23. Test 23 contains both filtering and noise. Test 45 contains some filtering but is dominated by VoIP variable delay; test 50 is a critical test containing noise and muting distortions, but no filtering. The mean and worst-case correlation over the whole database of 45 subjective tests are also given. To assist with comparison with PESQ, Table 4.4 shows the correlation for each model minus the corresponding correlation coefficient for PESQ (E).

Table 4.3: Frequency response equalisation and model performance

	Individual tests				All tests	
Model	**16**	**23**	**45**	**50**	**Mean**	**Worst-case**
(A) PSQM, ITU-T P.861	0.6148	0.6172	0.3082	0.8729	0.7775	0.2792
(B) PESQ, no frequency response equalisation	0.6223	0.7950	0.9154	0.8531	0.8779	0.6095
(C) PAMS 3.1	0.9409	0.9029	0.9402	0.7645	0.9296	0.7645
(D) PESQ, standard spectral difference	0.9281	0.9026	0.9061	0.8109	0.9434	0.8109
(E) PESQ, ITU-T P.862	0.9275	0.9028	0.9059	0.8108	0.9435	0.8108
(F) PESQ, smoothed phaseless cross-spectrum	0.9355	0.8906	0.9096	0.8091	0.944	0.8091
(G) As (F) using coherence weighting with with λ=1.	0.9353	0.8963	0.9077	0.8114	0.9442	0.8114
(H) As (F) using coherence weighting with with λ=2.	0.9340	0.9007	0.9082	0.8129	0.9442	0.8129
(I) As (H) but using smoothed spectral difference	0.9288	0.9045	0.9132	0.8108	0.9439	0.8108

The weakest model is PSQM (A). The low correlation on tests 16 and 23 is due to the lack of transfer function equalisation. Test 45 includes variable delay, which PSQM does not model.

Disabling the frequency response equalisation process in PESQ (B) shows its influence on the model's accuracy, which drops by 6.5% on average and by 30.5% for test 16. However, it appears that in two variable-delay tests, 45 and 50, removing the frequency response equalisation significantly improves the performance of PESQ; both of these tests contain severe muting distortions.

PAMS (C) uses a very different auditory transform and performs transfer function equalisation and variable delay processing in a different way from PESQ. This is why it performs better than PESQ (E) for tests 16 and 45. However, PAMS does not model rapid gain variation and is more susceptible to the muting problem, which give it lower accuracy than PESQ for test 50 and overall.

Table 4.4: Frequency response equalisation, performance compared to PESQ

	Individual tests				All tests	
Model	**16**	**23**	**45**	**50**	**Mean**	**Worst-case**
(A) PSQM, ITU-T P.861	-0.3126	-0.2856	-0.5978	0.0621	-0.1661	-0.6369
(B) PESQ, no frequency response equalisation	-0.3052	-0.1078	0.0095	0.0423	-0.0656	-0.3052
(C) PAMS 3.1	0.0134	0.0001	0.0343	-0.0463	-0.0139	-0.1280
(D) PESQ, standard spectral difference	0.0006	-0.0002	0.0001	0.0001	-0.0001	-0.0019
(E) PESQ, ITU-T P.862	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
(F) PESQ, smoothed phaseless cross-spectrum	0.008	-0.0122	0.0036	-0.0018	0.0005	-0.0122
(G) As (F) using coherence weighting with with λ=1.	0.0078	-0.0065	0.0018	0.0006	0.0007	-0.0065
(H) As (F) using coherence weighting with with λ=2.	0.0065	-0.0021	0.0022	0.0021	0.0007	-0.0044
(I) As (H) but using smoothed spectral difference	0.0013	0.0017	0.0073	-0.0000	0.0004	-0.0024

The only difference between Model (D) and PESQ (E) is the removal of the ±20dB gain limit. This has little effect on the overall quality score because the subjective tests used here do not include extreme frequency responses such as severe bandlimiting, notch filters or resonance peaks. PESQ (E) is likely to be more accurate in these cases due to the limitation on equalisation.

Models (F)–(I) show the effect of alternative perceptual transfer function estimation methods in PESQ. The phaseless cross-spectrum method (F) improves accuracy for clean speech (test 16) but reduces it for noise. The worst-case performance of this method is improved using the local coherence weighting with λ=2 (H), suggesting that the problem with test 23 may be due to estimation bias. Using smoothed spectral difference (I) with the same coherence weighting gives a consistent, if limited, improvement compared to PESQ (E).

4.6 Conclusions

Linear filtering is common in telephone networks, and is likely to be encountered in end-to-end measurements using analogue or acoustic interfaces. The frequency responses of a number of telephone terminal devices, and the models used to simulate these responses in subjective testing, were introduced.

Filtering using these devices appears to have little subjective effect with PCM coders. However, signals with heavy low-frequency content, which should ideally be controlled by the send filter, can adversely affect the performance of low bit-rate speech coders.

Without any process to estimate or equalise linear transfer functions, PSQM gives very inaccurate predictions for conditions that include filtering. This is because it treats linear and non-linear distortions in the same way, leading to filtered conditions being given much poorer quality scores than unfiltered conditions with the same subjective MOS.

Several techniques are available for linear transfer function estimation. As there is typically little knowledge of the structure of the system under test, it may be difficult to determine the appropriate model order for parametric methods. Two nonparametric methods were also discussed. Cross-spectrum-based TFE provides good noise rejection but is biased by non-linear effects such as clock jitter. Spectral difference is very resilient to the non-linear distortions considered here, but is biased by noise at low SNR.

It was shown that the performance of PSQM with linear filtering can be greatly improved by estimating the transfer function magnitude, and equalising the reference signal to the degraded signal to compensate for filtering in the system under test. Transfer function estimation and equalisation may be performed using the auditory transform of a perceptual model.

Cross-spectrum estimation may be performed effectively in the perceptual domain by discarding the phase component. While this means that the frequency response estimate is no longer asymptotically unbiased, it substantially reduces the bias due to non-linear processes that was found with the complex linear method. The phaseless cross-spectrum method also provides 3–10dB better noise rejection than spectral difference.

Perceptual smoothing may be applied to limit the equalisation performed for strong bandlimiting. Local frame coherence can reduce estimation bias due to periods of severe distortion such as noise or muting, and was found to improve the noise rejection of the spectral difference method by giving higher weight to speech frames.

Using these methods, the highest overall performance was found by modifying PESQ to use phaseless TFE with perceptual smoothing and local coherence weighting; however, the improvement over Hekstra's simpler spectral difference method is marginal. The results clearly show that perceptual transfer function equalisation allows PAMS and PESQ to give much more accurate predictions than PSQM for subjective tests that include filtering.

CHAPTER 5

Multi-parameter regression for perceptual quality assessment

5.1 Overview

While the previous two chapters have focused on how to make perceptual models robust to delay and linear filtering, the accuracy and generality of a model is strongly influenced by the output stage, which computes a quality score from a number of distortion parameters, and by the data that is used to train and test this. Following other authors, the term "cognitive model" will be used in this thesis to describe this stage of the model; of course in practice this is far from a model of general human perception or cognition.

The problem of developing a cognitive model is characterised by the following. The relationship between perceptual distortion and MOS is in general non-linear, with threshold and compression effects in the perception of loudness and of audible differences, and in their averaging over time. Subjective quality is known to be a multi-dimensional problem, with the potential for quite different cognitive treatment of distinct classes of error such as noise, deletion, linear filtering or pitch variation. There is significant, but unknown, variability on the target output (subjective MOS); these deviations are both systematic, particularly between subjective tests, and random or pseudo-random in the subjects' choice of vote and the influence of order and material. Finally, the data is expensive. A commercial P.800 listening quality subjective test using 24 subjects costs US$15–25,000, and can only practically measure about 50 conditions because of constraints on listening time, limiting the number of factors that can be explored in any one test.

The multi-dimensional nature of subjective quality was the focus of the diagnostic acceptability measure (DAM), which was introduced by Voiers and popularised by Quackenbush et al. [Voiers 1977, Quackenbush 1988]. This asks subjects to rate quality on a number of different properties, such as "fluttering/bubbling" for the speech signal, or "rumbling/thumping" for the background noise. DAM also maps these subjective parameters to a number of overall measures: total speech quality, total background quality, acceptability, intelligibility and pleasantness.

A consequence of measuring so many subjective quality parameters is that they may lead to conflicting results. For example, noise reduction systems typically show an improvement in background quality but may be neutral in speech quality and give reduction in intelligibility. Whilst for specific design applications this level of analysis may be of interest, it makes it difficult to draw conclusions on overall quality. For this reason, it is often preferable to focus on a single subjective quality scale, requiring the subjects to use their own judgement and experience to combine the many different types of distortions that they hear to a single score. The ACR listening quality scale [ITU-T P.800, ITU-T P.830] shown in Table 1.1 has become by far the most common for telecoms, so that, at the time of writing, MOS can normally be assumed to mean ACR LQ MOS.

Before the start of this work, most perceptual models focused on extracting a single distortion parameter that, when averaged over time and frequency, correlates well with MOS. The internal coefficients of the model, such as the loudness transform, signal scaling/equalisation, and asymmetry weighting, were tuned to maximise correlation for this distortion parameter. A good example of this is PSQM [ITU-T P.861], which was introduced in chapter 2. PSQM also clearly showed the limitation of the use of a single distortion parameter in the effect of background noise; effectively, by using the silent interval weighting, distortions during speech and noise were treated separately.

The concept of computing an overall quality score from multiple parameters derived using objective measures was considered by Quackenbush in the 1980s [Quackenbush 1988], and continued to be used by authors including Voran [Voran 1999a]. However, these were all essentially non-perceptual models based on simple objective measures or spectral distance, and these authors only considered linear mappings. PAMS [Hollier 1995] and PEAQ [ITU-R BS.1387] were amongst the first perceptual models to specifically measure multiple distortion parameters and map these to quality score in a final cognitive process. Both the parameters and the mapping process differ substantially between models. Hollier proposed the use of error activity and error entropy, and their combination using a non-linear function in a manual training process [Hollier 1995]. In PEAQ, which was introduced in section 2.5.5, 5 or 11 parameters were computed to describe different error classes, and these were mapped to an objective difference grade using a neural network [ITU-R BS.1387].

Several authors have used neural networks for the cognitive model. The MLP network in PEAQ has 36 free coefficients, and was trained with a few hundred effective data points. Tarraf and Meyers used SNR in perceptual bands as input to a family of three-layer MLP networks, one per talker, with five outputs to classify the score as *excellent...bad* [Tarraf 1999]. The author estimates that they had about a thousand free coefficients in total, trained using 90 different conditions. Radial basis functions were considered by Meky and Saadawi for use with

cepstral distance parameters, although they give little information on the network structure and order [Meky 1997].

The risk of over-training, which was also discussed in 2.5.5, is illustrated by the results of [Tarraf 1999]. Without cross-validation, their network reached a correlation of 0.997 for this small data set, but training with cross-validation (testing the model on data not used for training) gave a more plausible correlation of 0.929. This problem is not restricted to neural networks: it can occur wherever there are large numbers of parameters or free coefficients. The cross-validation approach [Bishop 1995] will be used in this chapter to address this, in particular for order selection.

A further problem is the systematic variability of subjective test data, which was introduced in chapter 2. Wang fitted a quadratic function for each subjective test to compare BSD, and other objective quality measures, to MOS [Wang 1992]. There is no guarantee that this will be monotonic – for three of the models that she evaluated, it was not – giving the models a higher correlation than would be the case if rank order was maintained. Such a mapping cannot be used in practice because one objective quality value may correspond to two or more different subjective scores. Other authors have used the logistic function, either for normalising MOS to an "objective" scale such as the MNRU [GSM HR3 1993] or to map objective speech quality to MOS [Hollier 1995, ITU-T P.861]. This is guaranteed to be monotonic, but has only a constrained range of curvature, which limits its usefulness. The author developed a method, described in section 5.3.4, for polynomial regression based on gradient descent using the multi-dimensional simplex method with a brick-wall cost function to enforce a monotonic constraint. This method was adopted by the ITU for the performance assessment of perceptual models for the P.862 competition [ITU-T P.862], and is currently in use for P.AAM and other developments [Beerends 2003].

The variation between tests is also important for model training. If it is not taken into account, there is a tendency to select parameters that predict this variability at the expense of accuracy at predicting MOS in general. This is shown in section 5.4.7 to be a significant problem, particularly as the subjective test database used for this work was much larger and more diverse than that considered by previous studies. The author used the method of monotonic polynomial regression to develop a generalised approach for model training that addresses both the problem of variability between tests and the robustness of the multi-dimensional fit. In addition, he was the first to describe how parameter selection can be integrated into this training process [Rix 1998a].

This chapter begins with a general formulation of the problem of perceptual model training, along with a description of the perceptual model that was used to produce the distortion parameter set used for this chapter and the subsets of the subjective test data that were used for test and training. Section 5.3 sets out three methods for mapping between objective and

subjective quality scores, for MOS normalisation and for performance assessment: the logistic function, linear regression, and monotonic polynomial regression. Following the procedure adopted by the ITU, the monotonic polynomial regression method has been used for performance assessment in this thesis.

The remainder of this chapter develops the multi-parameter model stage by stage, with a summary of performance results for each development. Section 5.4 discusses linear and non-linear regression methods that are applied directly, without experiment normalisation, to prediction of MOS. Section 5.4.7 describes two different approaches for normalisation to reduce the effect of systematic variations between experiments, and shows how these can be used in regression. Section 5.6 introduces four parameter selection techniques which may be used in conjunction with the regression and normalisation processes to improve the ability of the model to generalise. The training process is completed by the use of joint optimisation of the model parameters and experiment fits, which is described in section 5.7.

The performance achieved by these methods shows the accuracy of the overall perceptual model, including the time alignment and transfer function equalisation processes described in the previous chapters. A full set of results on a large database of subjective tests is given in section 5.8. Further details of the subjective test database may be found in Appendix D.

5.2 Problem formulation

5.2.1 Distortion perception

The following examples describe some types of distortion that are of interest for speech quality assessment. The error surface $e_s(k,f)$ is defined as the auditory transform of the degraded signal minus the auditory transform of the reference, for frame k and frequency band f.

Additive noise. Subjects recognise that noise, whether stationary or time-varying, is separate from the speech, even though it may mask parts of speech. If the noise level is low, it is usually only noticeable during silent periods. In a perceptual model, noise appears as additive errors. During silent intervals, with clean references, the error is entirely due to the noise, but the compression in the loudness transform means that the error becomes small where the speech is louder than the noise. If the loudness transform is correct, the partial loudness of the noise should be identical to the error.

Coding errors. Speech and audio coders are designed to minimise the perceived error. A good coder will give minimal perceived distortion; for a perceptual model, this will appear as positive or negative distortions that are close to zero in $e_s(k,f)$ due to the compression in the

loudness transform and masking. At lower bit-rates distortion may be more audible, and the range of error types depends on the coder and signal.

Muting. In conditions of packet loss or severe channel errors, systems may be unable to receive speech and usually play out silence. Front-end and back-end clipping may be caused by VAD and DTX, muting the start or end of speech utterances respectively. This type of distortion corresponds in the perceptual model to a sustained period of negative errors on $e_s(k,f)$ during speech.

5.2.2 Distortion parameter extraction

In order to compute a quality score, it is necessary to reduce the high-dimension error surface to a set of parameters that describe different modes of distortion. In BSD, Wang summed squared error within each frame and computed the mean over the duration of the signal [Wang 1992]:

$$BSD = \frac{1}{N_k} \sum_k \sum_f e_s(k,f)^2 \tag{5-1}$$

Other authors have considered taking the general L_p-norm rather than the mean squared error, or computing only the positive or negative part of $e_s(k,f)$ within each frame to give measures of positive and negative errors [Quackenbush 1988, Schroeder 1991]. A generalisation of this for two-stage temporal averaging is used in PESQ and P.AAM, and was discussed in section 2.5.4.

Hollier introduced error entropy, which had originally been suggested in the context of video coding, to compute a dimensionless measure of the distribution of errors [Hollier 1995]:

$$E_e = -\frac{1}{N_k} \sum_k \sum_f a(k,f) \log a(k,f) \tag{5-2}$$

where $a(k,f) = e_s(k,f) / \frac{1}{N_k} \sum_{k'} \sum_{f'} | e_s(k',f') |$

Beerends proposed that the error be weighted according to the relative power of the signals, to model the asymmetry effect [Beerends 1994, ITU-T P.861], and used a lower threshold or deadzone, to compute the PSQM noise disturbance $N(k)$ for frame k:

$$N(k) = \sum_f \max\left(0, |e_s(k,f)| - 0.01 \left(\frac{d_p(k,f)+1}{r_p(k,f)+1} \right)^{0.2} \right) \tag{5-3}$$

Here $r_p(k,f)$ and $d_p(k,f)$ are the reference and degraded signal excitation in units of power.

For this chapter, a development version of the P.AAM model was used to generate the distortion parameters. The author is collaborating on P.AAM with Beerends, Berger and

Goldstein. The parameters generated here are based on total error disturbance, with a modified deadzone, computed with and without an asymmetry weighting similar to that shown in (5-3), in a very similar process to that used to generate the two distortion parameters computed in PESQ (Appendix C). For this study, different parameters are calculated for speech only, silent periods only, and the whole signals, and by taking the absolute value, the positive part or the negative part only. A number of cases that do not contain useful information – for example, negative errors during silent periods – were pruned to avoid generating redundant parameters. In addition, the averaging described in section 2.5.4 is used with several different L_p powers for integration over frequency and the two stages of temporal integration. This gave a total of N_x=248 candidate distortion parameters, generated from 18 basic parameter structures.

5.2.3 Functional form

A set of distortion parameters $x(1)\cdots x(N_x)$ are computed by the perceptual model for each test case. Some subset $m_1\cdots m_M$ of M of these parameters is selected. Objective speech quality (OSQ) y_0 is predicted from this subset, using a function $f(\cdot)$ and the associated coefficients $\mathbf{a}$, as shown in equation (5-4). Unlike conventional regression, a further mapping is performed with coefficients $\mathbf{b}_s$ that are specific to each subjective test s, as shown in (5-5).

$$y_0 = f(x(m_1),\cdots x(m_M),\mathbf{a}) \tag{5-4}$$

$$\hat{y} = g(y_0,\mathbf{b}_s) \tag{5-5}$$

The prediction error associated with this test case is given by (5-6).

$$\varepsilon_{file} = \hat{y} - y = g(f(x(m_1),\cdots x(m_M),\mathbf{a}),\mathbf{b}_s) - y \tag{5-6}$$

As discussed above, in many cases we are interested in the average quality computed per condition, which consists of several individual test cases, as shown in (5-7). Thus the linear average of the error for each condition k in a test is used as the basic prediction error. (In the remainder of this chapter, k is used to denote a test condition, rather than the frame index used above.)

$$\varepsilon_{cond}(k) = \frac{1}{N_i(k)}\sum_i \hat{y}(i) - y(i) \tag{5-7}$$

To train a model, a cost or lack of fit function is required. Correlation coefficient leaves redundancy in the subjective test mappings, as gradient and scale are eliminated by Pearson's formula for the correlation coefficient (2-3). A convenient cost function which avoids this redundancy is to use a weighted mean squared error, as shown in (5-8). In this equation, the cost is summed for a number of subjective tests s, which may be separated between training

and test sets. Within each subjective test s, squared error is summed for each condition k. Note however that a condition may consist of multiple test cases, as given in (5-7).

$$C = \frac{1}{N_s} \sum_s \frac{w(s)}{Nk(s)} \sum_k \varepsilon_{cond}^2(k) \tag{5-8}$$

By comparison with equation (2-4), if the weights chosen are inversely proportional to the variance of each subjective test, minimisation of (5-8) is equivalent to maximising the mean square of the correlation coefficient calculated per condition, as shown in (5-9). Note that this only holds if (5-8) has been minimised, in which case (5-10) follows.

$$w(s) = \frac{1}{\sigma_y^2(s)} \qquad \frac{1}{N_s} \sum_s \rho^2(s) = \frac{1}{N_s} \sum_s \left[1 - \frac{\sigma_\varepsilon^2(s)}{\sigma_y^2(s)}\right] = 1 - \frac{1}{N_s} \sum_s \frac{\sigma_\varepsilon^2(s)}{\sigma_y^2(s)} = 1 - C \tag{5-9}$$

$$\sigma_\varepsilon^2(s) = \frac{1}{N_k(s)} \sum_k \varepsilon_{cond}^2(k) \tag{5-10}$$

For the case of a single subjective test (N_s=1), the mean squared correlation coefficient (5-9) is identical to the R^2 statistic in the literature on multiple regression (e.g. [Quackenbush 1988]).

5.2.4 Problem size

Table 5.1 gives the size of the data set and problem. The key figures are K, K_{eff} and N_x, the total number of conditions, the effective total number of conditions, and the total number of parameters respectively. This is the total number of data points available for model training and testing. In practice there is redundancy in the database: most subjective tests include standard references such as MNRU and ITU/ETSI standard coders, and many of the subjective tests used for this study had been conducted in two or more laboratories using essentially identical network conditions (Appendix D). This may reduce the effect of "noise" in voting errors or systematic variations between these cases, but is likely to result in clustering of the distortion parameters for each case. Thus the author has conservatively estimated the effective number of unique conditions K_{eff} as one quarter of the total K.

Table 5.1: Problem size

Symbol	Value	Description
N_s	45	Total number of subjective tests
$N_k(s)$	22–72	Number of conditions in a subjective test
$N_i(k)$	1–48	Number of files per condition
K	2119	Total number of conditions
K_{eff}	530	Estimate of effective number of conditions
	25648	Total number of files
N_x	248	Total number of distortion parameters

Because the total number of distortion parameters N_x is of the same order of magnitude as K_{eff}, it is clear that over-training is very likely to result if all parameters were used.

5.2.5 Test and training sets

To reduce the risk of over-training, and to allow model order to be evaluated, a cross-validation approach will be used in this chapter. The 45 subjective tests are split into a group of 15 tests used for training, termed the training set (T), and 30 tests used to validate the performance and generalisation, termed the validation set (V). The overall set of tests (T+V) will also be used for some training and assessments. The set that each test was assigned to is noted in Appendix D.

For most of the methods considered here, the parameters are selected and the model trained on set T only, without reference to set V. However, for the MLP, V is used to select the best of a family of models generated from set T, so is not strictly unknown. For further evaluation of the MLP than considered here, it would be advisable to split the validation set to keep some data completely unknown; this allows a more critical test of a model's ability to generalise [Bishop 1995, Demuth 2001].

5.3 Variability of subjective MOS

Section 2.6.4 introduced the variability of MOS between subjective tests conducted using the ACR method [ITU-T P.800]. This section considers two approaches to regression or normalisation between subjective and objective quality: the logistic function, which was used by the ITU until the late 1990s, and a monotonic polynomial method that was proposed by the author. Example results show the relative performance of these approaches for assessing the accuracy of PESQ and PSQM.

5.3.1 Logistic function

The sigmoid logistic function provides a mapping that is guaranteed to be monotonic and to produce a bounded output. The generalised form of the logistic that has been used for speech quality measurement is given in equation (5-11) [ITU-T P.861, Voran 1999b].

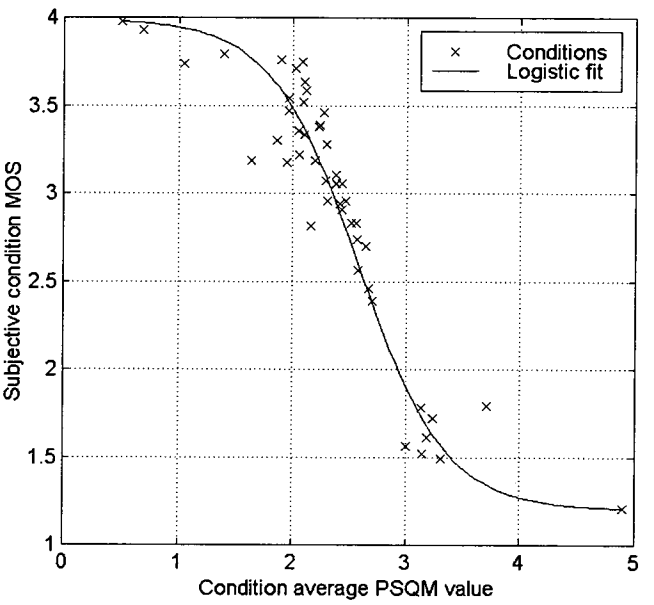
$$\hat{y} = a + \frac{b-a}{\exp(cy_0 + d)} \tag{5-11}$$

The output of the logistic is bounded in (a, b), and the relationship is increasing if c is negative. It is only invertible within (a, b), and in practice for output values close to these asymptotes, the input becomes large.

Two methods may be used to fit the four coefficients, depending on whether (*a*, *b*) are pre-determined or are optimised. Firstly, (*a*, *b*) are often set to (1, 5), as these are the limits of the MOS scale. Alternative pragmatic approaches have also been used, for example setting (*a*, *b*) to the lowest and highest MOS values, although this makes the function not invertible at these points. For values of *y* within (*a*, *b*), linear regression of $cy_0 + d$ against $\ln\frac{b-a}{y-a}$ can be used to find (*c*, *d*), although the resultant mapping is in general not minimum squared error between *y* and $\hat{y}$.

The second method is to perform general non-linear gradient descent to optimise the coefficients for minimum MSE between *y* and $\hat{y}$. Because the error is computed in this domain, invertibility is not required for regression. An example fit using this method to map from PSQM score to MOS is shown in Figure 5.1. For this example, the linear correlation, per condition, between PSQM and MOS is –0.8756. After mapping through the logistic it is 0.9574. Clearly in this case the non-linear mapping provides a much more flattering view of the performance of the model.

Figure 5.1: Example logistic fit



If invertibility is needed for other applications such as normalisation of *y*, this can be enforced by a cost penalty to constrain (*a*, *b*) to lie outside the range of *y*. Regression may also be performed in the input space with an error given by $\ln\frac{b-a}{y-a} - (cy_0 + d)$, but in this case invertibility is required and must be enforced if *a* and *b* are to be included in the optimisation.

The logistic method has commonly been used in perceptual modelling for two purposes. In the first case, for a known objective measure such as MNRU Q [ITU-T P.810], the inverse logistic was used to map MOS scores $y(k)$ to the equivalent Q value. This could be used to perform comparisons in a domain that was independent of the systematic variations between subjective tests [GSM HR3 1993]. Similarly, the forward logistic of equation (5-11) was used to map an objective distortion measure $y_0(k)$ such as PAMS score or PSQM value to an estimate of MOS $y(k)$ [Hollier 1995, ITU-T P.861].

By varying the bounds of the logistic, it can generalise between a straight line and a step function, but it cannot curve the other way. To compare two subjective tests using the logistic therefore requires an intermediate mapping to a domain such as Q, and this process is invalid if other factors cause a systematic offset in Q between tests. As discussed in section 5.5.1, this is a major problem, because the MNRU is influenced by the spectral content of the speech signals used in the test. This makes the comparison unreliable even for subjective tests of exactly the same design. Certain experimenters also prefer to omit the MNRU references if the other conditions in the test cover a sufficiently wide range, as the additional five or six conditions can allow more factors of interest to be evaluated.

Note that the logistic behaves like a step function for $c \to \infty$. It falls rapidly towards a very low gradient away from $cy_0 + d = 0$, as can be seen at the extremes of Figure 5.1. A phenomenon that was observed by the author is that a flat part of the logistic curve is assigned to sections at the top or bottom of the range of OSQ where the perceptual model has no information, substantially improving correlation in the presence of outliers. This is particularly common for PSQM in regions where the model gives very inaccurate scores. For a critical evaluation of a model's performance, this is undesirable, for the following reasons.

Firstly, without access to the subjective test data required to produce a mapping such as that derived in Figure 5.1, users can only compute the basic OSQ score. It is misleading if a large difference (say, 0.5 PSQM value) that is observed between two conditions may be compressed to near zero at the ends of the range, but remain large in the centre of the range, as a result of the logistic function. A user could conclude that a highly audible difference exists when in fact it does not. If a compressive function is necessary for good correlation with MOS, it should be applied as part of the OSQ computation rather than in the performance analysis.

Secondly, it has been found by the author that, particularly with PSQM, strong flattening is only used in some subjective tests. It seems implausible that there should be information in the measure in the extremes of the OSQ scale for some tests, but not in others. An assessment method should not allow the mapping function gradient to go to zero in this way to avoid this discrepancy.

5.3.2 Polynomial regression

If the relationship between a perceptual model score $y_0(k)$ and MOS is not sigmoid, the logistic function may reduce the accuracy of that perceptual model that is estimated using measures such as correlation coefficient and RMSE. The author therefore developed an alternative process, to use polynomials that are guaranteed to be monotonic in a given range. The cubic polynomial derived using this method is now most the common mapping used by ITU-T study group 12 for evaluation of perceptual models. This has the same number of coefficients as the logistic, but much more freedom in its curvature, so it can be used for a direct comparison between subjective tests or from perceptual models to subjective MOS. In addition, the polynomial with low order is unable to flatten completely in any part of the range in the way observed with the logistic, so it does not suppress outliers as strongly.

The general polynomial form, including a constant offset, is shown in equation (5-12).

$$\hat{y} = b_m y_0^m + b_{m-1} y_0^{m-1} + \cdots b_1 y_0 + b_0 \tag{5-12}$$

For K data points, $K>m$, the optimum fit between $y(k)$ and $\hat{y}(k)$ in a minimum MSE sense is found from (5-13), which can be derived using the method of least squares:

$$\mathbf{b} = (\mathbf{Y}_0^T \mathbf{Y}_0)^{-1} \mathbf{Y}_0^T \mathbf{y} \tag{5-13}$$

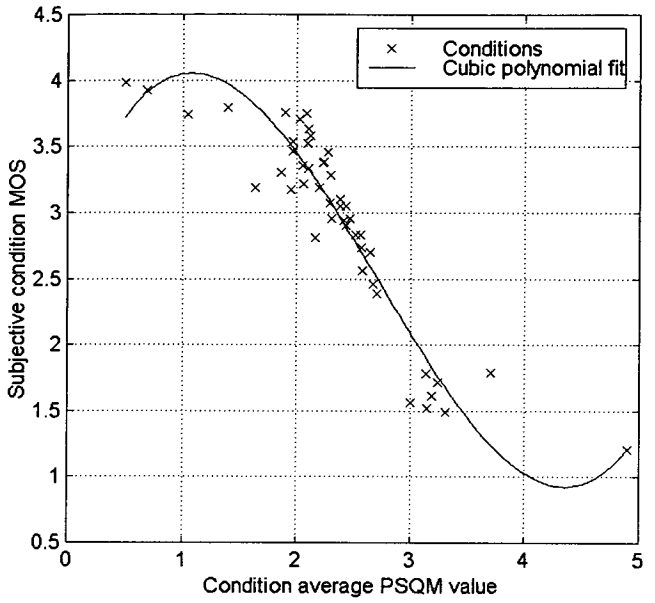
where

$$\mathbf{b} = \begin{bmatrix} b_m \\ \vdots \\ b_0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y(1) \\ \vdots \\ y(K) \end{bmatrix} \quad \mathbf{Y}_0 = \begin{bmatrix} y_0^m(1) & \cdots & 1 \\ \vdots & & \vdots \\ y_0^m(K) & \cdots & 1 \end{bmatrix} \tag{5-14}$$

and the matrix $\mathbf{Y}_0$ has at least m+1 linearly independent rows. Alternative methods may be used if this is not the case and the fit is underdetermined. For $K \leq m$, the prediction error is zero at the training points $\mathbf{y}$.

An example of the application of this method to PSQM is shown in Figure 5.2, showing the same data as Figure 5.1. The linear correlation, per condition, between PSQM and MOS is –0.8756 in this subjective test. After mapping through the cubic polynomial the correlation is 0.9429.

Figure 5.2: Example polynomial fit



5.3.3 Monotonic constraint

The example shown in Figure 5.2 illustrates the main problem with using polynomial regression: there is no guarantee that the mapping is monotonic. In Figure 5.2, as PSQM value drops below 1.0 (corresponding to a reduction in degradation), the estimated MOS goes down, when it would be expected to increase. Inversion effects such as this are common with subjective test data, as shown in Table 5.2, and make the results essentially meaningless. The author therefore developed monotonically constrained polynomials for this application, using the method described in the next section.

The author also proposed to generalise this concept to the whole output stage of the perceptual model, as given by equations (5-4) and (5-5). The motivation for this is as follows. The auditory transform and distortion parameter calculation can be arranged so that, for each class of degradation, each distortion parameter either increases with the amount of degradation or remains constant. Any set of training data can only include a limited set of modes of these parameters and this may lead to parameters being given weights of opposite sign – effectively trading them off against each other. The author found this to be common, particularly where there is non-linearity in the data or the parameters and the model is over-determined. This is the case in the published models MNB [Voran 1999, ITU-T P.861] and PEAQ [ITU-R BS.1387]. If in future another distortion type is encountered that has a different correlation or mode between the parameters, the result may be an inversion in the relationship, making OSQ rise with increasing degradation. Preserving a monotonic relationship between each parameter and

OSQ in the cognitive model is likely to avoid this and increase model robustness in training [Rix 1998a].

5.3.4 Monotonic polynomials

Since polynomials of order 2 or more may have turning points, it is necessary to establish some range to ensure that the polynomial is monotonic in this range. The input value (OSQ) must if necessary be bounded to ensure that it remains within this range. Some perceptual models have a pre-determined, bounded output range: for PAMS, the author constrained the output to [1, 5]; for PSQM, it is bounded to [0, 6.5]. Alternatively, for a given subjective test, the bounds may be set at the lowest and highest values of OSQ. For the following, it is assumed that $v_1 \le y_0 \le v_2$ and it is required that the polynomial should be monotonically non-decreasing. (The extension to the case of monotonically decreasing polynomials is trivial.)

The requirement for the polynomial to be monotonically non-decreasing in $[v_1, v_2]$ is equivalent to requiring that the gradient is non-negative, as shown in (5-15). For polynomials up to order 4, it is sufficient to evaluate this requirement at v_1, v_2, and at any inflection points y_i (where the gradient is a maximum or a minimum) in this range. Inflection points y_i are satisfied by (5-16), which requires finding the real roots of an order m–2 polynomial. Further checks are needed for order 5 and above to deal with coincidence of inflection and turning points, but for the purposes of this thesis we will restrict ourselves to order 4 or below.

$$\frac{d\hat{y}}{dy_0} = mb_m y_0^{m-1} + (m-1)b_{m-1}y_0^{m-2} + \cdots b_1 \ge 0, \quad v_1 \le y_0 \le v_2 \tag{5-15}$$

$$\frac{d^2\hat{y}}{dy_i^2} = m(m-1)b_m y_i^{m-2} + (m-1)(m-2)b_{m-1}y_i^{m-3} + \cdots 2b_2 = 0 \tag{5-16}$$

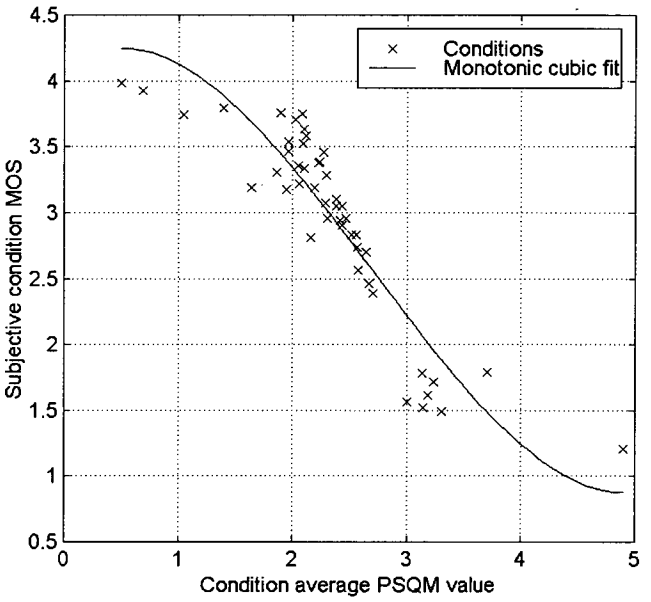
For a quadratic polynomial, it is sufficient to evaluate the gradient (5-15) at the bounds $[v_1, v_2]$. For a cubic, the gradient must be checked at $[v_1, v_2]$ and at one potential inflection point at $-b_2/3b_3$, if it lies within this range. For a quartic, the solution of (5-16) gives a quadratic equation; if $q = 36b_3^2 - 96b_4b_2 \ge 0$, the gradient must be checked at the roots $(-6b_3 \pm \sqrt{q})/24b_4$, if they lie in the range, as well as at the bounds.

The author implemented this method to find the polynomial fit with minimum MSE subject to the monotonic constraint. A cost function was used that returns infinity if the candidate fit fails any of the tests described above, or the MSE $\frac{1}{N_k}\sum_k (\tilde{y}(k) - y(k))^2$ otherwise. An initial fit is found using a candidate polynomial that satisfies the monotonic constraints. If the fit for the required order m is not already monotonic, a monotonic fit of lower order is found and a

standard gradient descent algorithm is used to optimise this for minimum MSE using this cost function. The Nelder-Mead multi-dimensional simplex search, implemented as the fmins tool in Matlab [Lagarias 1998, Mathworks 1998], was found to be suitable for this purpose. This algorithm is robust to discontinuities in the cost function such as the constraint used here; it is able to contract to reach an optimum that is very close to a constraint boundary. To reduce the risk of failure due to the simplex stalling [Lagarias 1998], the fmins procedure was invoked twice.

The result of applying this algorithm to PSQM for the same data as described in the previous sections is shown in Figure 5.3. Note that PSQM value increases with the amount of distortion, so the fit chosen here is monotonically decreasing. The bounds on PSQM values chosen were the minimum and maximum PSQM values for this subjective test. As before, the linear correlation, per condition, between PSQM and MOS is –0.8756 in this subjective test. After mapping through the monotonic cubic polynomial the correlation is 0.9221, considerably lower than the 0.9429 for the unconstrained cubic because the gradient is only allowed to go to zero at the bounds of the range.

Figure 5.3: Example monotonic polynomial fit



A weakness of the monotonic polynomial method is that it may produce a fit that is not invertible for all MOS $y(k)$ in a test. There are several ways to address this. The cost function method described above may be extended to require that the output values at the bounds $[\upsilon_1, \upsilon_2]$ include the required range of $y(k)$. Alternatively, for cubic polynomials, the Bezier spline formulation [Bronshtein 1985] can be arranged to provide a convenient way to control the output

range and ensure that the function is monotonic. Assuming that the input, conventionally denoted t, is in the range [0,1] (this can be mapped from $[\upsilon_1, \upsilon_2]$ by substitution), this formulation can be transformed directly to and from the polynomial coefficient vector **b** using **b** = **M** **p**, where the invertible matrix **M** and Bezier coefficient vector **p** are given by

$$\mathbf{M} = \begin{bmatrix} -1 & 3 & -3 & 1 \\ 3 & -6 & 3 & 0 \\ -3 & 3 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} \tag{5-17}$$

The coefficients p_1 and p_4 set the lower and upper bounds corresponding to t=0 and t=1 respectively. The polynomial on t is guaranteed to be monotonically increasing in (0,1) provided that $p_1 \le p_2 \le p_4$, and $p_1 \le p_3 \le p_4$. This provides a simple way to evaluate the output range and ensure that the function is monotonic.

5.3.5 Results

The performance of PESQ and PSQM was measured by computing the per experiment mappings described in this section over the subjective test database, and comparing them to the linear fit. The logistic function used allowed all parameters $a...d$ to be optimised to achieve the minimum MSE between $y(k)$ and $\hat{y}(k)$. The average and worst-case absolute correlation coefficient for the models with each of the mapping methods is presented in Table 5.2.

Table 5.2: Effect of mapping functions on correlation coefficient

Model	Mapping method	Mean correlation	Worst-case correlation	Monotonic?
PESQ	Linear	0.9238	0.7960	Yes
	Logistic	0.9457	0.8056	Yes
	Cubic	0.9449	0.8108	No in 17 of 45 cases
	Monotonic cubic	0.9435	0.8108	Yes
PAMS	Linear	0.9154	0.6608	Yes
	Logistic	0.9316	0.7821	Yes
	Cubic	0.9309	0.7711	No in 17 of 45 cases
	Monotonic cubic	0.9296	0.7645	Yes
PSQM	Linear	0.7314	0.2558	Yes
	Logistic	0.7857	0.2961	Yes
	Cubic	0.7843	0.2795	No in 20 of 45 cases
	Monotonic cubic	0.7737	0.2792	Yes

This data illustrates that the extra freedom in all of the other mapping functions improves the performance metric compared to the linear mapping. The logistic, which allows the mapping to take on a strong bounded shape, generally gives the highest correlation. In some tests, as in the example shown in Figure 5.1, the logistic is very close to flat for a large part of the range of objective quality.

The worst-case performance with PESQ shows that the logistic is not always as "good" as the polynomial methods, giving lower correlation scores than the polynomial methods in some cases. This a consequence of the fact that the logistic function has fewer modes of curvature than the polynomials.

The monotonic constraint to preserve ordering is necessary if a polynomial method is to be used for performance evaluation: in more than a third of tests for each model, the unconstrained cubic fit was not monotonic, giving an unrealistically high correlation. The monotonic polynomial regression process described above addresses this problem.

An advantage of the monotonic polynomial method is that it can only reach zero gradient at points in the range, typically only at the ends. Compared to the logistic function, in which the gradient may be near zero for large parts of the range, this means that the monotonic polynomial method provides a more critical assessment of the information content of the objective measure throughout the output range, as discussed in section 5.3.1.

5.4 Regression methods

The cost function proposed in (5-8) will normally be non-linear in the coefficients $\mathbf{a}$ and $\mathbf{b}_s$. However, the problem may be simplified either by ignoring variation between experiments or by some invertible normalisation process. The latter will be discussed in the next section. Under this simplification, conventional techniques for regression or function approximation may be applied.

For the purposes of this section, y_0 (5-4) will be trained directly against y for the training set of experiments introduced in section 5.2.5, and performance will be evaluated using monotonic polynomial regression per experiment. Five basic regression methods are considered: linear regression, Volterra non-linear regression, sigmoid multi-layer perceptron, and two extensions of the monotonic polynomial method, either applied individually to each parameter or as a general non-linear mapping. In each case the methods are applied to three different parameter sets, and comparative results are presented in section 5.4.6.

5.4.1 Linear regression

The linear model for predicting speech quality using M distortion parameters is shown in equation (5-18). This is an application of the method of least squares introduced above.

$$y_0(k) = a_0 + a_1 x_1(k) + \cdots a_M x_M(k) \tag{5-18}$$

In general, for N_k data points, $N_k > M$, the optimum fit between $y(k)$ and $y_0(k)$ in a minimum MSE sense is given by (5-19):

$$\mathbf{a} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{5-19}$$

where

$$\mathbf{a} = \begin{bmatrix} a_0 \\ \vdots \\ a_N \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y(1) \\ \vdots \\ y(N_k) \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1(1) & \cdots & x_M(1) \\ \vdots & \vdots & & \vdots \\ 1 & x_1(N_k) & \cdots & x_M(N_k) \end{bmatrix} \tag{5-20}$$

and the matrix $\mathbf{X}$ has at least $N+1$ linearly independent rows. Alternative methods may be used if this is not the case and the fit is underdetermined, but this is not of interest for this thesis as it is very likely to lead to over-training of a cognitive model.

5.4.2 Volterra non-linear regression

Linear regression can be generalised to model non-linear functions by expanding the parameter set to include powers, cross-products and higher powers of cross-products, analogous to a Taylor series expansion of a multi-dimensional function. This has the advantage that the fit is still linear in the parameters, allowing rapid direct solution. In practice the number of parameters rises rapidly, particularly with products, so parameter selection methods as discussed in section 5.6 are normally also required. For this section, only the parameters, the squares, and the first-order products will be considered, extending $\mathbf{X}$ to the matrix with $(M^2+3M)/2+1$ columns shown in (5-21). The minimum MSE solution is given by (5-19).

$$\mathbf{X} = \begin{bmatrix} 1 & x_1(1) & \cdots & x_M(1) & x_1^2(1) & \cdots & x_M^2(1) & x_1(1)\cdot x_2(1) & \cdots & x_{M-1}(1)\cdot x_M(1) \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & x_1(N_k) & \cdots & x_M(N_k) & x_1^2(N_k) & \cdots & x_M^2(N_k) & x_1(N_k)\cdot x_2(N_k) & \cdots & x_{M-1}(N_k)\cdot x_M(N_k) \end{bmatrix}$$

$$\mathbf{a} = \begin{bmatrix} a_0 \\ \vdots \\ a_{(M^2+3M)/2} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y(1) \\ \vdots \\ y(N_k) \end{bmatrix} \tag{5-21}$$

5.4.3 Sigmoid multi-layer perceptron

The neural network approach has been applied to many problems of pattern recognition and is considered here as a general non-linear function approximation method [Bishop 1995]. One of the most common network structures is the multi-layer perceptron (MLP) with one or more hidden layers. For this evaluation, following the cognitive model of PEAQ [ITU-R BS.1387], one hidden layer was used and the activation function was the sigmoid, a simplified version of the logistic function of (5-11), as shown in (5-22).

$$\mathrm{sig}(x) = \frac{1}{1+\exp(-x)} \tag{5-22}$$

Each node of the network computes a linear weighted sum, including an offset, and processes this through the activation function. In common with many authors including [ITU-R BS.1387], a linear mapping is applied to the input parameters so that each parameter is given roughly equal overall weight. With one hidden layer containing H nodes, and M input parameters, the network function is given by (5-23).

$$y_0 = \gamma_1 + \gamma_2 \,\mathrm{sig}\left\{ b_0 + \sum_{h=1}^{H} b_h \,\mathrm{sig}\left(a_{h,0} + \sum_{n=1}^{N} a_{h,n} \frac{x_n - \alpha_n}{\beta_n - \alpha_n} \right) \right\} \tag{5-23}$$

This model has $H(M+1)$ input weights $a_{h,n}$ and $H+3$ weights b_h, γ_1, γ_2 in the output node. The linear input mapping of each parameter is performed to re-scale the range of x_n from $[\alpha_n, \beta_n]$ to [0,1], where α_n and β_n are the minimum and maximum values of x_n in the training data (note that the parameters are not actually bounded to this range). A final linear output mapping, defined by γ_1, γ_2, is required for the analysis used in this section because the output data (MOS) is not within the [0,1] output range of the sigmoid.

Three key problems with MLPs for this function approximation problem are as follows. (i) The gradient of sig(x) falls exponentially towards zero for |x| >> 0. This can give problems with some gradient descent algorithms. (ii) Particularly for small data sets, it can help prevent roughness and aid generalisation to unseen data if network training is stopped before it has reached a global minimum. This method, termed validation or cross-validation, will be used here. (iii) Order selection, in particular the number of hidden layers and the number of nodes in each, is not straightforward and is often determined heuristically by training and testing networks of different structures, although some theoretical methods have been developed to treat this in a more rigorous way [Bishop 1995].

In common with many regression techniques, roughness or inversion effects may arise with over-training of MLPs. To address this for the cognitive model application, the author developed a method [Rix 1999c] to apply the concept of a monotonic input-output relationship (that was introduced in section 5.3.3) to the MLP, by replacing the weights with their values passed

through a function that is always positive. A useful function is shown in (5-24), which is the integral of sig(x). pwt(x) tends to x for $x >> 0$, and to exp(x) for $x << 0$. A weight function was chosen rather than a brick-wall cost constraint as many gradient descent algorithms may have difficulties with the large number of discontinuities in the cost surface that would result.

$$\text{pwt}(x) = \log(\exp(x) + 1) \tag{5-24}$$

The positive weight function must not be applied to the offsets $a_{h,0}$ and b_0, as these may need to be negative and do not affect whether the network is monotonic. With this positive weight function, the outputs of all nodes in the network are monotonically non-decreasing with respect to all input parameters. The final sign of the relationship is determined by γ_2, but it is not usually necessary to constrain this in training as it is determined by the data. An advantage of this approach is that the network can be used to model a monotonic function of either increasing or decreasing sign. The modified monotonic MLP with one hidden layer using this method is shown in (5-25).

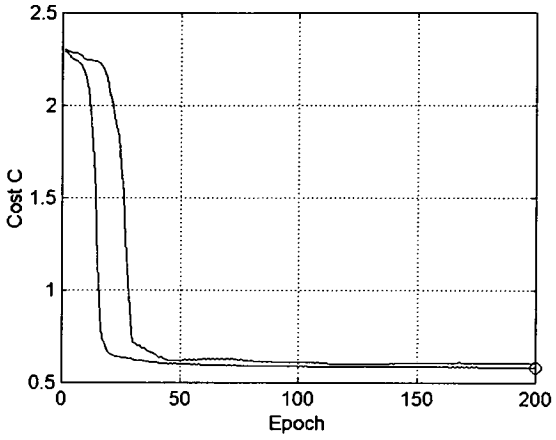
$$y_0 = \gamma_1 + \gamma_2 \,\text{sig}\left\{ b_0 + \sum_{h=1}^{H} \text{pwt}(b_h)\,\text{sig}\left(a_{h,0} + \sum_{n=1}^{N} \text{pwt}(a_{h,n}) \frac{x_n - \alpha_n}{\beta_n - \alpha_n} \right) \right\} \tag{5-25}$$

For the evaluation of MLPs for the cognitive model, MLPs of the form of equations (5-23) and (5-25) were both tried using several different values of H and with multiple different initialisations. The Nelder-Mead multi-dimensional simplex search was used as a simple gradient descent algorithm for training. Although this method is not particularly efficient, it was chosen (i) because the same search algorithm was also used for the polynomial methods described in this chapter, and (ii) because the simplex method takes bounded steps and adaptively expands or contracts the step (simplex) size depending on the data, which are desirable behaviours for training MLPs.

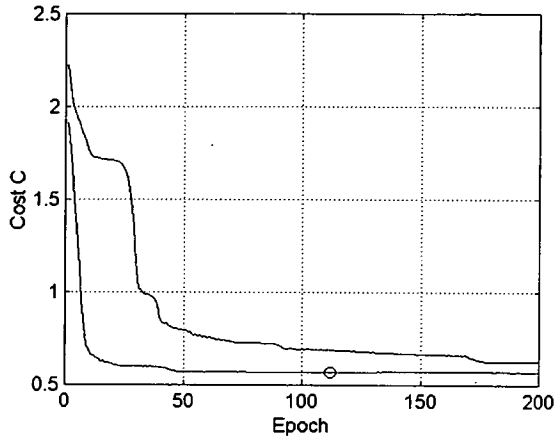
The results of applying this training process to predict condition MOS for the 4-parameter training set (see section 5.4.6) with H=8 and 49 free coefficients are shown in Figure 5.4(a) for the standard MLP (5-23) and in Figure 5.4(b) for the constrained MLP (5-25). Ten different initialisations were used in each case and the data is plotted every "epoch" i.e. 698 iterations of the gradient descent algorithm. Both graphs show the lower and upper bounds on cost measured on the combined training and validation set (T+V). These figures indicate that the monotonic constraint on the MLP increases the variance on cost but does not necessarily make the resultant fit any worse, or make the regression slower, for this dataset.

Figure 5.4: Results of training MLP and constrained MLP

(a) MLP. Best cost 0.581 on combined set (T+V) at epoch 200.



(b) Monotonic constrained MLP. Best cost 0.566 on combined set (T+V) at epoch 112.



For the results presented here and in section 5.4.6, gradient descent was performed using the test data set (T) only, with MSE as the cost function. Several values of H were tried, and ten random initialisations were used in each case. The optimum model was chosen based on minimum MSE calculated for both the test (T) and validation (V) data sets from any of the initialisations, again computed every 698 iterations (there are 698 conditions in the data used for training).

5.4.4 Non-linear parameter normalisation

If the relationship between individual parameters and the output is non-linear, linear regression methods in particular may give solutions that are far from optimal when evaluated using a non-linear mapping such as that described in section 5.3.4. Conversely, non-linear methods such as the MLP require a slow iterative search with a large number of free coefficients. The author found that a simple approach which normalises (linearises) and bounds each parameter prior to linear regression could produce improved results whilst still permitting direct regression. Unlike Volterra regression, the monotonic constraint is straightforward to implement in this approach.

This process consists of the following steps, applied to each parameter in turn.

(1) Find some bounds $[v_1, v_2]$ on the parameter, and bound the parameter using these values.

(2) Perform monotonic polynomial regression to find an optimal fit from the bounded parameter to MOS for the chosen order, using the algorithm described in section 5.3.4.

(3) Map the parameter through this polynomial.

The mapped parameters are then used in place of the original parameter set for standard linear regression, as described in section 5.4.1.

For the purposes of this section, $[\upsilon_1, \upsilon_2]$ were found from the values 0.1% and 99.9% through the sorted list of each parameter in the training data, and polynomial order 3 was used. Because the optimum choice of bounds and order is in practice data-dependent, this method is better suited to the creation of large numbers of candidate parameters or as an initialisation for the joint search method described in the next section.

5.4.5 Joint multi-parameter monotonic polynomial regression

In practice there is little reason to assume that the optimum non-linear relationship between any one parameter and MOS will give the best overall solution in linear regression, in particular when combined with a final non-linear mapping such as that used for performance assessment. For the training of PAMS, the author explored a more general approach: to jointly optimise the non-linear mappings on all of the chosen parameters.

This leads to the formulation of the cognitive model shown in equation (5-26), in this case using cubic polynomials.

$$y_0(k) = a_0 + a_{1,3}x_1^3(k) + a_{1,2}x_1^2(k) + a_{1,1}x_1(k) + \cdots a_{M,3}x_M^3(k) + a_{M,2}x_M^2(k) + a_{M,1}x_M(k) \quad (5\text{-}26)$$

Initialisation is performed using the method described in section 5.4.4, with parameters that are inverted treated separately. The full set of $3N+1$ weights $a_{n,p}$ are optimised by gradient descent using the simplex search (section 5.3.4), with a cost function that returns the MSE if all of the individual polynomials with coefficients $a_{n,3}$, $a_{n,2}$ and $a_{n,1}$ are monotonic, or infinity otherwise.

The author subsequently extended this method to a joint optimisation of the cognitive model and the parameter set, which is discussed in section 5.7.

5.4.6 Results

The results for this section are intended to be illustrative, as in practice some form of optimum parameter selection and per-experiment normalisation are required. This is explored further in the next two sections.

The parameter sets shown in Table 5.3 were used to evaluate the performance of each regression method. These parameter sets were selected as optimum for direct linear regression against polynomial normalised MOS using McHenry's selection method (section 5.6.5).

Table 5.3: Parameter sets used for direct regression

Set	Number of parameters M	Parameters
a	2	P.AAM asymmetric error, whole signal, p=1.0, 8.0, 4.0 P.AAM disturbance total, whole signal, p=3.5, 4.0, 2.0
b	4	P.AAM asymmetric error, whole signal, p=1.0, 8.0, 3.0 P.AAM disturbance total, whole signal, p=3.0, 3.0, 3.0 P.AAM disturbance total, whole signal, p=3.5, 3.0, 2.0 P.862 positive disturbance in silence, muting option 1, p=3.0, 2.0, 2.0
c	8	P.AAM disturbance total, whole signal, p=2.5, 4.0, 2.0 P.AAM asymmetric error, whole signal, p=1.0, 8.0, 4.0 P.AAM asymmetric error, speech, p=1.0, 8.0, 4.0 P.AAM asymmetric error, speech, p=1.5, 2.0, 4.0 P.862 disturbance total, whole signal, muting option 0, p=2.0, 4.0, 2.0 P.862 positive disturbance in silence, muting option 0, p=2.0, 4.0, 2.0 P.862 disturbance total, speech, muting option 1, p=2.0, 4.0, 2.0 P.862 disturbance total, speech, muting option 1, p=3.0, 2.0, 2.0

The accuracy of each method is assessed by computing the normalised MSE C (5-8), with each subjective test given weight inversely proportion to its variance using $w(s) = \frac{1}{\sigma_y^2(s)}$. This means that C=1 corresponds to no predictive power, while C=0 would correspond to perfect prediction. Table 5.4 shows the results for each of the parameter sets listed in Table 5.3. Each regression method is applied to the training set (T), and then the cost for the validation set (V) is calculated separately; the methods are also applied to training and testing on the whole data set (T+V). The tables also show the number of coefficients in the fit, the computation time to perform the regression once for the given set of parameters, and whether the fit on the training data set is monotonically decreasing between each parameter and y_0. The time is averaged over 100 runs for the linear, Volterra and normalised parameter methods. One-off pre-processing such as the computation of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{y}$ (5-19) or the non-linear normalisation of the parameters is not included, because it is the time to evaluate a candidate parameter set that is of greatest interest for parameter selection.

Table 5.4 also presents the mean and worst-case correlation coefficient ϱ, calculated using (2-3), per condition, with monotonic cubic mapping per experiment.

Table 5.4: Results for direct regression against condition MOS

(a) 2 parameters

Method	Order	Time, s	Mono.	C (T)	C (V)	C (T+V)	ϱ_{mean} (T+V)	ϱ_{min} (T+V)
Linear	3	0.0003	yes	0.616	0.540	0.565	0.9407	0.8426
Volterra	6	0.00411	no	0.572	0.590	0.584	0.9399	0.8100
NormBnd	3	0.0011	yes	0.628	0.542	0.571	0.9391	0.8394
MultPoly	7	1.45	yes	0.570	0.540	0.550	0.9422	0.8424
MLP	19 (*H*=4)	159	no	0.545	0.556	0.552	0.9407	0.8273
Monotonic MLP	19 (*H*=4)	161	yes	0.548	0.550	0.549	0.9405	0.8253

(a) 4 parameters

Method	Order	Time, s	Mono.	C (T)	C (V)	C (T+V)	ϱ_{mean} (T+V)	ϱ_{min} (T+V)
Linear	5	0.0004	no	0.595	0.532	0.553	0.9409	0.8530
Volterra	15	0.0114	no	0.523	0.646	0.606	0.9350	0.7593
NormBnd	5	0.0016	no	0.611	0.541	0.564	0.9397	0.8491
MultPoly	13	2.3	yes	0.563	0.553	0.556	0.9399	0.8542
MLP	51 (*H*=8)	425	no	0.612	0.567	0.582	0.9437	0.8402
Monotonic MLP	51 (*H*=8)	443	yes	0.533	0.583	0.566	0.9372	0.8169

(a) 8 parameters

Method	Order	Time, s	Mono.	C (T)	C (V)	C (T+V)	ϱ_{mean} (T+V)	ϱ_{min} (T+V)
Linear	9	0.0004	no	0.577	0.539	0.552	0.9420	0.8680
Volterra	45	0.054	no	0.463	0.610	0.561	0.9300	0.8241
NormBnd	9	0.0025	no	0.609	0.538	0.561	0.9410	0.8552
MultPoly	25	4.19	yes	0.588	0.522	0.544	0.9376	0.8246
MLP	43 (*H*=4)	573	no	0.663	0.565	0.597	0.9408	0.8283
Monotonic MLP	43 (*H*=4)	581	yes	0.576	0.606	0.597	0.9340	0.8430

5.4.7 Discussion

The most striking point about these results is the poor normalised cost *C* for all of the regression methods. This is a consequence of the high variation in underlying MOS, which the model cannot predict, and means that equation (5-9) cannot be applied to compute the mean squared

correlation coefficient from C. Because the variability between subjective tests is not generally of interest for perceptual models, as discussed in section 2.6.4, the correlation coefficients are a more faithful guide to the model's performance in this case.

For these parameters, the linear mapping method is the most consistent, giving the highest, or close to the highest, ϱ_{mean} and ϱ_{min} for all three orders, and the lowest cost C on the unknown validation data for 2 and 4 parameters. The best single model is the linear fit with 8 parameters, which has ϱ_{mean}=0.9420 and ϱ_{min}=0.8680 respectively.

5.5 Normalisation–based training

The variability of MOS between subjective tests poses a significant problem for model training. The author has found that at least ten subjective tests must be used to include enough factors to describe the complexity of current telephone networks. If inter-test variability is not taken into account, there is a tendency to select parameters that predict this variability at the expense of accuracy at predicting MOS in general.

Two alternative methods for removing this variability by normalisation of MOS are investigated in this section. The starting point is to use some objective quality estimate to compute the inverse of the subjective test mapping function (5-5), for each subjective test, and to apply this to compute a normalised MOS y' according to (5-27).

$$y' = g^{-1}(y, \mathbf{b}_s) \tag{5-27}$$

Section 5.5.3 gives results for normalisation performed as a one-off pre-processing of MOS prior to the application of the regression techniques described in section 5.4. The use of normalisation for iterative training is considered in section 5.7.

5.5.1 Normalisation by MNRU conditions

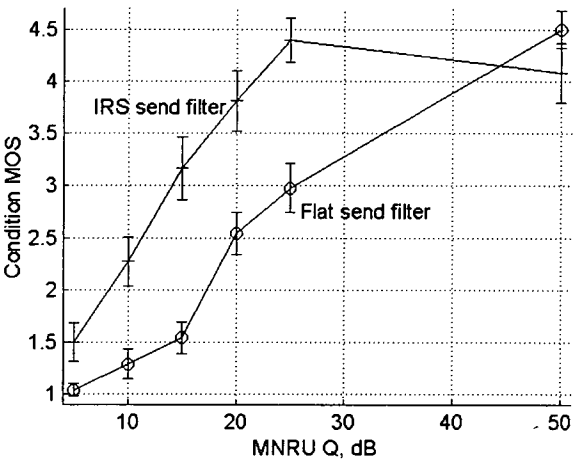
The method of normalisation to MNRU equivalent Q was introduced above in section 5.3.1. This has been used to map subjective MOS through the inverse of the logistic function given in equation (5-11), to compute a dB-like distortion measure [ITU-T P.810, GSM HR 1993, ITU-T P.861]. For the normalisation considered here, the mapping was trained on the MNRU conditions in each subjective test, and the inverse mapping used to compute equivalent Q values for each condition. An average forward logistic mapping was then used to map the equivalent Q values to normalised MOS for all conditions.

To address the problem that the MNRU is not invertible for $y \notin (a, b)$, conditions at or above these limits were set to a very low (0) or large (100) value of Q. Condition average y was used;

the averaging process reduces the number of cases that exactly reach 1 or 5, and reduces the spreading effect of the strong curvature of the logistic near the asymptotes.

One limitation of MNRU normalisation is that the relationship between *Q* and MOS varies with the spectrum of the input speech, as was observed in [GSM HR3 1993]. This is clearly illustrated in Figure 5.5, which shows results, with estimated 95% confidence intervals, for the MNRU conditions in subjective test 16, which included two types of send filter (see also Figure 4.6 and Figure 4.23). The IRS send filter serves to de-emphasise low frequencies, which contain most of the energy in speech, and boosts the level in the 2–3kHz region which contributes strongly to intelligibility. This results in a flattening of the speech spectrum, and means that the speech-modulated white noise that is added by the MNRU is effectively about 10dB lower, compared to the speech spectrum at these frequencies, than for the flat filtered speech.

Figure 5.5: Relationship between *Q* and MOS



Because of this effect, the results for applying the MNRU normalisation method cannot be compared between different forms of pre-processing of the speech material. However, several different types of recording microphone and pre-filtering are available for subjective testing, and the exact processing used is not known in many cases. For this reason, only the 24 subjective tests where the send filter is known to conform to the IRS or MIRS send characteristic [ITU-T P.830] were used for the evaluation in section 5.5.3.

5.5.2 Normalisation to candidate objective score

The weakness of MNRU for normalisation can be addressed by using a more accurate objective measure such as an existing perceptual model, or a model developed using the direct regression process described in the previous section. Since performance will be assessed using the monotonic 3rd-order polynomial method described in section 5.3.4, the process

chosen was to fit a polynomial to map OSQ to MOS, ensuring that it is invertible for all values of MOS in the given test. The polynomial can then be inverted using Cardano's method or a numerical root-finding technique.

The algorithm described in section 5.3.4 is not suitable for this purpose as it does not guarantee that the polynomial will be invertible for the range of y in the test. However the formulation introduced in section 5.3.4 in terms of Bezier splines can be used, as two of the Bezier coefficients are the maximum and minimum values for inputs at the bounds. It is sufficient to ensure that these lie outside the maximum and minimum values of y for the given subjective test. This was implemented for non-linear gradient descent by transforming the corresponding coefficients through the pwt() function shown in (5-24) and adding or subtracting the result from the required limiting value of y.

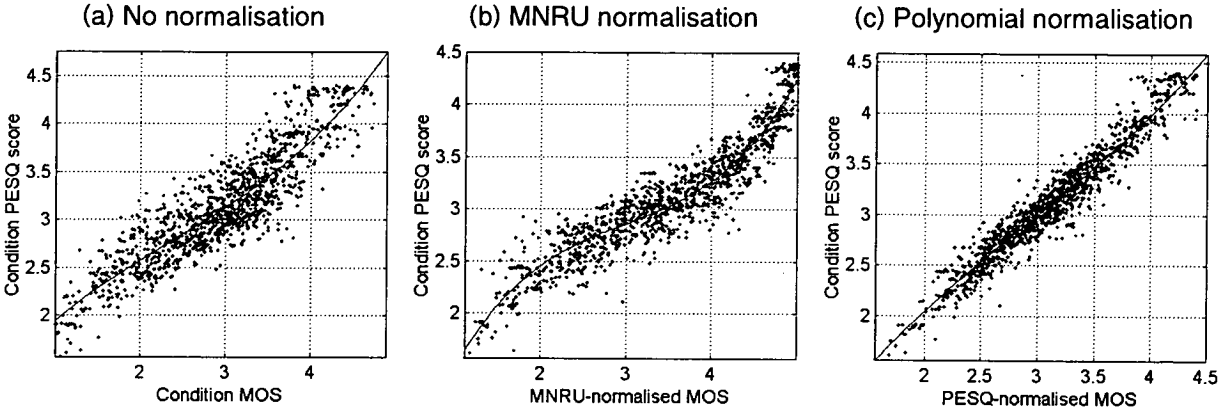
The optimum fit was found in the minimum MSE sense between $\hat{y} = g(y_0, \mathbf{b}_s)$ and y for each subjective test, calculated per condition, using P.862 PESQ score as $\hat{y}$. One-dimensional function minimisation was then applied to compute normalised MOS y' according to (5-27).

5.5.3 Results

The performance of the two different normalisation procedures was evaluated by testing the reduction in bias between MOS and PESQ score achieved by the normalisation. Because of the pre-filtering problem identified in section 5.5.1, the comparison between the methods could only be performed for tests where the input filtering was known to match the IRS or MIRS send filters. The polynomial normalisation method was also evaluated over the whole dataset.

The results of applying the two normalisation methods to the IRS/MIRS dataset are shown by the scatter plots in Figure 5.6, which show the raw, or normalised, condition MOS compared to condition PESQ score [ITU-T P.862], along with the monotonic cubic regression line computed using the method described in section 5.3.4. These graphs show that, while the MNRU normalisation method in Figure 5.6(b) appears to slightly reduce scatter compared to raw MOS (Figure 5.6(a)), it makes the relationship with PESQ score even more non-linear. The polynomial normalisation method shown in Figure 5.6(c) substantially reduces scatter and gives a regression line that is almost exactly linear, which is to be expected as the same objective model was used for normalisation and testing.

Figure 5.6: Comparison of MOS normalisation methods



The performance of the methods was also evaluated by calculating the correlation coefficient and RMSE, between PESQ score and MOS, over the IRS/MIRS subjective tests and over all subjective tests (for the raw MOS and the polynomial normalisation only). This was carried out both with a linear fit and with monotonic cubic regression, to evaluate how well the normalisation process had succeeded in linearising the problem. RMSE and correlation were calculated by linearising against the condition PESQ score, which is the same for each normalisation method. These results are shown in Table 5.5.

Table 5.5: Performance of MOS normalisation methods

(a) IRS/MIRS dataset (24 subjective tests)

	Linear		Cubic mapping	
Normalisation	**Correl.**	**RMSE**	**Correl**	**RMSE**
None (raw condition MOS)	0.8895	0.275	0.8944	0.250
MNRU	0.9209	0.289	0.9439	0.185
Polynomial with PESQ	0.9665	0.145	0.9671	0.142

(b) All subjective tests (Polynomial method only)

	Linear		Cubic mapping	
Normalisation	**Correl.**	**RMSE**	**Correl**	**RMSE**
None (raw condition MOS)	0.8716	0.299	0.8745	0.280
Polynomial with PESQ	0.9511	0.180	0.9513	0.178

It is clear from Table 5.5(a) that the MNRU normalisation makes a small improvement in correlation; however, it makes the relationship between PESQ score and MOS much more non-linear, which causes RMSE for the linear case to rise. In both Table 5.5(a) and Table 5.5(b), the invertible polynomial normalisation method gives a large improvement in correlation. The relationship is not exactly linear as a result of the requirement that the mapping can be inverted;

however, the fact that the difference in correlation coefficient between the linear and cubic methods is only 0.0002 indicates that the normalisation method is very close to linear.

Because the polynomial method reduces scatter much more effectively than the MNRU method, and has the further advantage that it can be applied to all subjective tests, the polynomial method was used to normalise MOS for direct regression in the remaining examples in this chapter.

Table 5.6 presents an evaluation of the performance of MOS normalisation using the monotonic cubic polynomial and PESQ score [ITU-T P.862]. The results are presented in the same way as for direct regression in Table 5.4. The parameter sets listed in Table 5.3 were used. See section 5.4.6 for a description of the methods used to produce these results and the values shown.

5.5.4 Discussion

The process of normalisation using MNRU appears to give little improvement compared to direct regression against MOS. This is likely to be a consequence of the problem of repeatability of the MNRU for different subjective test material.

However, normalisation using the polynomial method, shown in Table 5.6, gives an increase in model performance compared to regression against raw MOS (Table 5.4) in almost every case. The mean correlation coefficient has improved by around 0.5–1% with most methods, and the worst-case correlation by up to 2%. The best single model for the direct regression, the linear fit with 8 parameters, is also one of the best models here, with ϱ_{mean}=0.9486 and ϱ_{min}=0.8669 which compare with ϱ_{mean}=0.9420 and ϱ_{min}=0.8680 for the same parameters and method in direct linear regression: an improvement of 0.46% on the mean, but a reduction of 0.11% in the worst-case correlation.

Interestingly, the normalisation process allows some of the non-linear regression methods to generalise better than direct linear regression. This is thought to be because the reduction in variability in MOS means there is less noise, allowing the free coefficients to be used to improve accuracy and generalisation. However, none of these is able to improve both the mean and worst-case performance, and the risk of over-training means that the linear fit would, from this data, probably be preferred.

Table 5.6: Results for regression against polynomial normalised MOS

(a) 2 parameters

Method	Order	Time, s	Mono.	C (T)	C (V)	C (T+V)	ϱ_{mean} (T+V)	ϱ_{min} (T+V)
Linear	3	0.0003	yes	0.1133	0.1102	0.1112	0.9472	0.8435
Volterra	6	0.0025	no	0.1103	0.1129	0.1121	0.9469	0.8383
NormBnd	3	0.0013	yes	0.1155	0.1120	0.1132	0.9465	0.8391
MultPoly	7	1.42	yes	0.1081	0.1114	0.1103	0.9477	0.8361
MLP	11 (H=2)	93	no	0.1140	0.1126	0.1131	0.9471	0.8428
Monotonic MLP	11 (H=2)	92	yes	0.1140	0.1126	0.1131	0.9471	0.8428

(b) 4 parameters

Method	Order	Time, s	Mono.	C (T)	C (V)	C (T+V)	ϱ_{mean} (T+V)	ϱ_{min} (T+V)
Linear	5	0.0004	no	0.1082	0.1156	0.1132	0.9475	0.8582
Volterra	15	0.0117	no	0.1045	0.1182	0.1137	0.9486	0.8216
NormBnd	5	0.0018	no	0.1091	0.1149	0.1130	0.9477	0.8547
MultPoly	13	2.21	yes	0.1084	0.1156	0.1132	0.9475	0.8585
MLP	15 (H=2)	111	no	0.1091	0.1173	0.1146	0.9482	0.8587
Monotonic MLP	15 (H=2)	124	yes	0.1055	0.1162	0.1127	0.9469	0.8568

(c) 8 parameters

Method	Order	Time, s	Mono.	C (T)	C (V)	C (T+V)	ϱ_{mean} (T+V)	ϱ_{min} (T+V)
Linear	9	0.0004	no	0.0995	0.1181	0.1120	0.9486	0.8669
Volterra	45	0.0575	no	0.0886	0.1112	0.1038	0.9533	0.8542
NormBnd	9	0.0024	no	0.1049	0.1135	0.1107	0.9489	0.8671
MultPoly	25	3.88	yes	0.1136	0.1131	0.1133	0.9470	0.8382
MLP	43 (H=4)	493	no	0.1032	0.1105	0.1081	0.9501	0.8553
Monotonic MLP	43 (H=4)	491	yes	0.1127	0.1120	0.1123	0.9477	0.8387

5.6 Parameter selection methods

It was mentioned above that it may be desirable to extract a large number N_x of parameters and select some small optimum subset of size M. This section explores why this is necessary, and evaluates four methods of parameter selection for cognitive modelling. Most of these parameter selection methods are used to generate a candidate set, which is tested using full regression. One method considered, multivariate adaptive regression splines (MARS) [Friedman 1991], performs both parameter selection using a forward-backward search, and non-linear regression. It is not within the scope of this thesis to provide a full evaluation of subset selection techniques; this section aims to apply some standard methods to cognitive modelling.

5.6.1 Parameter selection in model training

The motivation for computing large numbers of parameters and selecting some optimum set is as follows. Data from experimental psychology can be used to derive functions and constants such as temporal and frequency resolution, loudness scaling, and masking. However, these data are usually gathered in basic experiments on low-level properties of the auditory system using synthetic signals such as tones or noise.

There is much less knowledge about the perceptibility of distortions, in particular where they are well above threshold, as is the case for this thesis, or with complex signals such as speech. The best function forms, and constants such as threshold, growth powers, and averaging powers, must be deduced empirically. On a 1GHz Pentium III computer, a typical perceptual model such as PESQ takes about 1s to process an 8s test file – or about 8 hours to process the entire database described in Appendix D. Most of this time is spent in filtering, time alignment, and the auditory transform. While iterative optimisation of model parameters is necessary to improve these components, it is very slow. In comparison, computing each perceptual distortion parameter takes less than 1ms. Provided that the other model components are held constant and the parameter selection process takes no more than a few days, it is therefore more efficient to generate many parameters with different combinations of coefficients, and to select the best, than to run the model repeatedly.

For the remainder of this chapter, selection will be performed from a large set of parameters that are produced using the methods described in section 5.2.2. A number of basic parameter structures (e.g. asymmetric error measured during speech) are used to generate multiple parameters with different values of p used in the three averaging stages.

Clearly this means that many parameters are derived from a given basic parameter structure. The main advantage of the ability to select two or more versions of one basic parameter is that this allows greater flexibility in the non-linear relationship between the underlying parameter and output score. Conversely, it is difficult to ensure that the relationship is monotonic in this case if

the selected versions can be given opposite signs, and in practice it was found by the author that "trade-offs" between similar parameters are common.

Principal component analysis (PCA) and related subspace transformations, which have been used to good effect in other applications for reducing the dimension of the parameter space, were only investigated briefly in this study for similar reasons: the high correlation between versions of each parameter combined with orthogonal PCA means that the main principal directions trade off many similar (but non-linearly and stochastically related) parameters. This makes it very difficult to avoid roughness in the resultant mapping and hence to keep the relationship monotonic.

As was the case for the results presented in the previous section, over-training is a significant risk if large numbers of parameters may be selected. The results given below are based on model training and separate testing using the data sets introduced in section 5.2.5, using PESQ-normalised MOS as the target quality score.

5.6.2 Exhaustive search

A way to guarantee that the optimum subset of M parameters is chosen is to exhaustively evaluate all possible combinations. This number of combinations is shown in equation (5-28), which, for $N_x >> M$ is $\Theta(N_x^M/M!)$.

$$\frac{N_x(N_x-1)\cdots(N_x-M+1)}{M!} = \frac{N_x!}{M!(N_x-M)!} \tag{5-28}$$

This number of combinations for N_x=248 is shown in Table 5.7, along with the approximate time required if it took 1ms to evaluate each combination. (In practice, the evaluation of a candidate set is typically $\Theta(M^3)$, but this rises much less quickly than (5-28).) Clearly this method is only really tractable for $M \leq 4$ unless N_x can be significantly reduced.

Table 5.7: Number of combinations in exhaustive search

M	1	2	3	4	5	6	7	8
Combinations	248	3.1×10^4	2.5×10^6	1.5×10^8	7.5×10^9	3.0×10^{11}	1.1×10^{13}	3.2×10^{14}
Approx. time	0.25s	31s	42 min	43 hours	87 days	9.6 years	330 years	10,000 years

As mentioned in the previous section, the parameter generation process used here creates many parameters from one basic structure. A practical assumption may be made that only one of each "family" of parameters may be selected, where each family is derived from one basic parameter structure, with variations in threshold and L_p powers, as discussed in section 5.2.2. With 18 families each containing an average of 14 parameters, the number of combinations is

reduced to the values shown in Table 5.8, again with the time required for 1ms per combination. This provides most benefit for large M, but in this case does not allow any more parameters to be explored. This was therefore not pursued further in this study.

Table 5.8: Number of combinations in reduced search

M	1	2	3	4	5	6	7	8
Combinations	248	3.0×10^4	2.2×10^6	1.2×10^8	4.6×10^9	1.4×10^{11}	3.4×10^{12}	6.5×10^{13}
Approx. time	0.25s	30s	37 min	33 hours	53 days	4.4 years	110 years	2,000 years

5.6.3 Forward or backward selection

An alternative process for parameter selection is to find the best single parameter, then iteratively find another parameter that gives the best performance in combination with this, and so on. This is known as forward selection, and is not guaranteed to find the optimum parameter set, as shown by the results below. Forward selection is much faster than exhaustive search, however, requiring $\Theta(MN_x)$ evaluations of the regression process for $N_x>>M$. For M=20 and N_x=192, this is under 4,000 combinations, making this process feasible for the most complex regression methods introduced in section 5.4.

Backward selection is the opposite of forward selection: the starting set consists of all available parameters, and parameters are progressively deleted until the desired size is reached. This requires about $N_x^2/2$ (i.e. $\Theta(N_x^2)$) evaluations of the regression process – 18,000 in this example. However, the fastest regression process, direct linear regression for M parameters, requires a matrix inversion to be performed which cannot normally be implemented in less than $\Theta(M^3)$ operations. Since the first cases have $M\approx N_x$, backward selection effectively uses $\Theta(N_x^5)$ computations – very much slower than forward selection, which requires $\Theta(M^4N_x)$ computations. Full backward selection was therefore not considered further.

5.6.4 Forward-backward selection and extensions

A common improvement of the forward selection method is to select a larger subset than is ultimately required, then progressively delete parameters using backward selection. A typical rule of thumb is to forward select $2M$ parameters, and then delete M of these using backward selection [Friedman 1991]. This requires about $2MN_x$ evaluations of the regression process – a factor of two more than forward selection – and is considered below.

This method was improved by performing multiple forward and backward selection for order up to M=8 until no more changes were made in the optimum sets.

In cases where evaluation of candidate parameter sets is very slow, even forward selection can be prohibitively complex. Several heuristic techniques can be applied to build parameter sets more quickly than forward selection, for example by combining the best pairs or triples of parameters; however, there is no guarantee that these will produce a solution that is better than forward selection. An alternative solution would be to use a fast method such as linear regression to select a small number of candidates, and then to optimise this using a slower non-linear regression method.

5.6.5 McHenry's method

An alternative method to improve a subset selected using forward regression, or some other non-optimal technique, was proposed by McHenry [McHenry 1978]. This paper described a method to improve the computational efficiency of selection of candidate parameters in the linear model using Wilks' Λ, which is not pursued here although it can be used, in particular, to speed up parameter selection for linear regression. The paper also proposed the following two-stage method to verify or improve a parameter subset.

The starting point is some candidate parameter subset of size M, which may for example be derived using forward selection. The first check iteratively tests replacing each parameter in the subset with each parameter that is not in the subset. If this results in a lower cost, the new subset is adopted and the check is re-started. The first check terminates when each parameter in the subset is found to be optimum, which is guaranteed because the cost is always non-increasing. The number of evaluations of the candidate parameter set depends on the data and the quality of the starting subset, but has been found by the author to be roughly $\Theta(MN_x)$ for a given order M.

The second check iteratively tests deleting each parameter and replacing it with the next-best alternative parameter that is not in the subset. This generates M new subsets. The first check is then repeated on each of these subsets, with the condition that the deleted parameter is not allowed to be re-selected. The entire process is re-started if a subset is identified that has lower cost than the initial candidate. This was found to require about $\Theta(M^2N_x)$ evaluations of the regression for a given order M.

A complete parameter selection process using this technique evaluates $\Theta(M^3N_x)$ candidate subsets. This makes it considerably more complex than forward selection but much cheaper than exhaustive search. The number of evaluations may be reduced to $\Theta(M^2N_x)$ if only the first check is performed, but this was found to lead to an increase in RMSE of about 2%.

While the full McHenry selection procedure is much faster than exhaustive search for $M>1$, it is not guaranteed to reach the same solution. In practice it was found that the full McHenry

selection method often finds the same optimum as exhaustive search for the low orders $M \leq 4$ for which the exhaustive method could be evaluated, and the sets found using McHenry's method are almost always better than forward, forward/backward and repeated forward/backward selection, as shown in the results below.

5.6.6 Multivariate adaptive regression splines

MARS is a general method that combines order selection, subset selection and non-linear regression. A commercial implementation of the algorithm based on the original work by Friedman was used; the following is an overview based on and using the notation of [Friedman 1991].

The first step in MARS is to create a set of (potentially non-linear) M_{max} basis functions that partition the parameter space, using forward selection. The process starts with a single constant basis function $B_1(\mathbf{x})=1$. At each iteration, each existing basis function $B_m(\mathbf{x})$ is considered for partitioning along all new dimensions v (i.e. where parameter x_v is not currently included in $B_m(\mathbf{x})$), at candidate knots t along x_v, taken from the data points which fall in the positive region of that basis function. The candidate partition creates two new basis functions $B_m(\mathbf{x})[+(x_v-t)]_+$ and $B_m(\mathbf{x})[-(x_v-t)]_+$, where $[x]_+=x$, $x>0$; $[x]_+=0$, $x \leq 0$. The optimum linear mapping on the basis functions is found for each candidate partition, and the two new basis functions from the best partition are added to the set.

The second step iteratively removes each basis function (apart from $B_1(\mathbf{x})$). At each stage the basis function whose deletion gives the best overall model is removed from the set. The optimum model in this sequence is considered to be the best candidate. The cost function for both steps 1 and 2 is based on the MSE of the linear mapping, weighted by a factor that penalises increasing number of parameters and knots. A limitation of the method is that the knot penalty weight d is unknown, although [Friedman 1991] suggests values in the range [2, 4], and this may affect the optimum model order M^* that is selected. It is advised that M_{max} is set to $2M^*$.

The model produced at this stage is continuous in the parameters but has step changes in gradient at each knot. This is improved by an optional post-processing step that fits a cubic function in the region of each knot, making the gradient continuous. The two boundary points that are used to determine this smoothing either side of each knot are placed mid-way between adjacent knots along a given parameter, so that adjacent knots share a boundary point. This minimises the number of coefficients introduced by this process. No further optimisation of knot locations is performed.

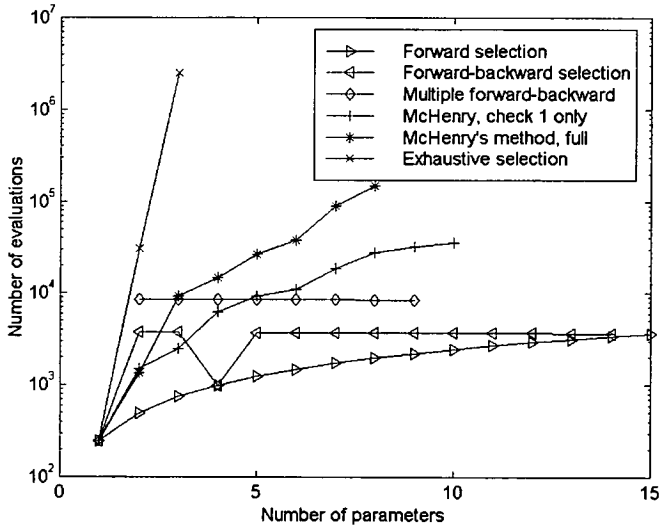
The partitioning process allows MARS to generalise from a purely linear mapping to one combining arbitrary products of parameters with a lower or upper threshold on each. In the

presence of correlation between the parameters, the model tends to produce large values at the edges of the data, for the following reason. While it is possible for MARS to produce S-shaped functions, this can only be achieved by duplicating basis functions so that, along a given parameter, a lower threshold function is followed by an upper threshold function with a higher threshold value; the difference between these two allows the model to generalise from a linear mapping to a hard S shape for that parameter. For this to be the case for an n–dimensional product of parameters requires 2^n basis functions, which is unlikely to be achieved in practice due to the penalty on high orders. This problem was addressed in a more practical way for this thesis by pre-processing the parameters to bound them using the data in the same way as the bounding process described in section 5.4.4. Although this does not fully resolve the problem of large output values at the edges of the data, it does limit the impact of parameters with long-tailed distributions.

5.6.7 Results

The relative computational cost of each parameter selection algorithm is illustrated by Figure 5.7. This shows the number of parameter sets evaluated by each algorithm using the linear regression method applied to MOS normalised using the monotonic polynomial method. No data was available on the complexity of the commercial MARS implementation that was used; however, it completed processing within a few seconds for order up to 15, suggesting that a forward-backward selection method as described in [Friedman 1991] is used.

Figure 5.7: Parameter selection, number of function evaluations



The model fits achieved by each parameter selection method are shown in Figure 5.8 and Figure 5.9. Figure 5.8 shows the cost of the fit on the training dataset (T). As expected, the

cost decreases with increasing order. McHenry's full selection method gives the lowest cost in most cases of the non-exhaustive methods, and finds a set that is only slightly above the best case for exhaustive search at M=3. Figure 5.9 plots the cost of applying the fit, calculated for the optimum parameters on the training dataset, to prediction of the validation dataset (V). This illustrates the ability of the model to generalise and shows that over-training is generally most serious for the parameter selections that fit best to the training dataset. Figure 5.10(a) plots the costs for applying McHenry's method on datasets T and V, as well as the cost for the model re-trained on the combined dataset (T+V). Both Figure 5.9 and Figure 5.10(a) indicate that, for the linear regression method on this data, 2 parameters produce a model that generalises best to the unseen data.

Figure 5.10(b) plots the training and validation cost for MARS, and also shows the costs when MARS was applied to the combined dataset (T+V). It is very clear that MARS does not generalise as well from T to V as the linear model, despite processing performed in MARS to regularise the fit. None of the basis functions chosen by MARS in either of these training processes included any parameter interactions: they all used either a single lower or upper bound.

Figure 5.8: Parameter selection, best sets, cost on training data (T)

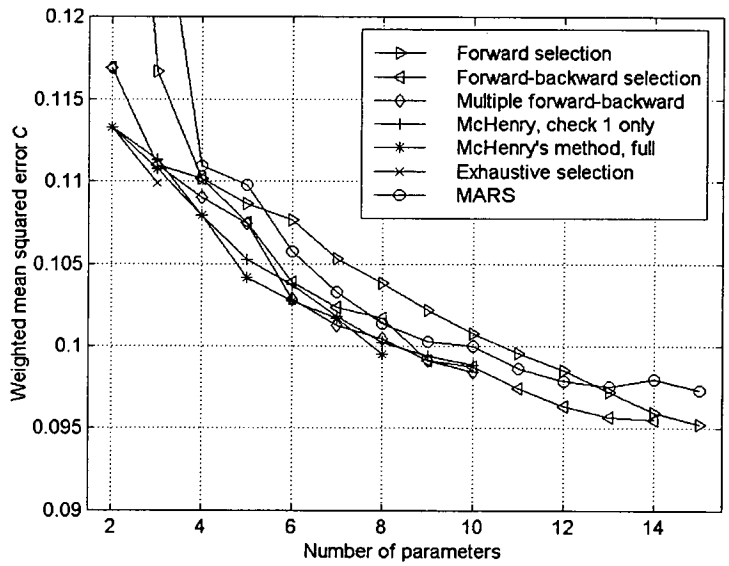


Figure 5.9: Parameter selection, best sets, cost on validation dataset (V)

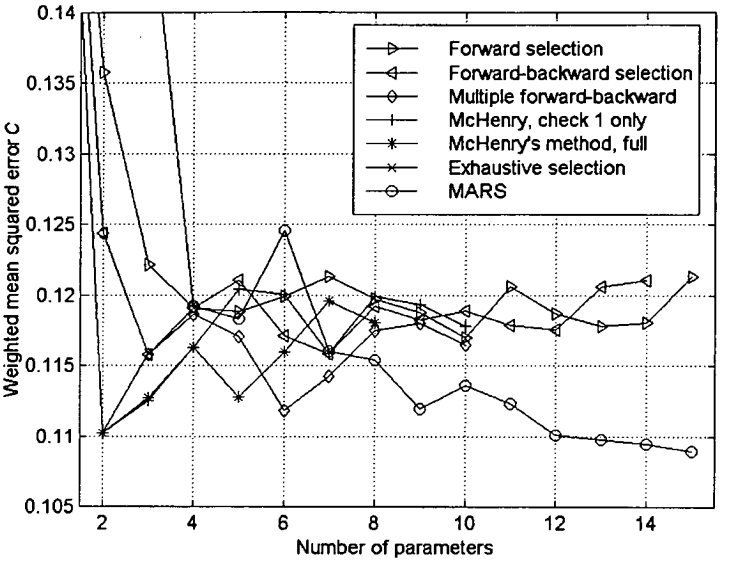
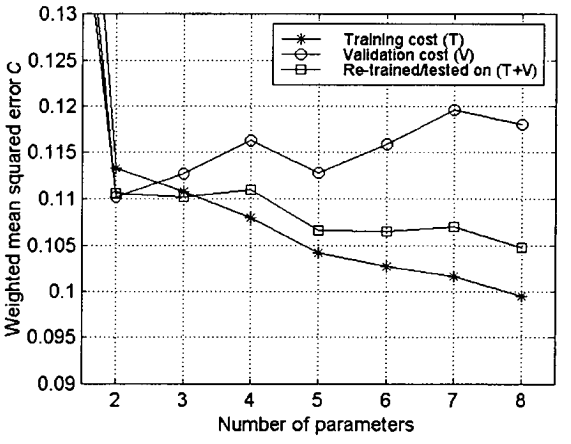
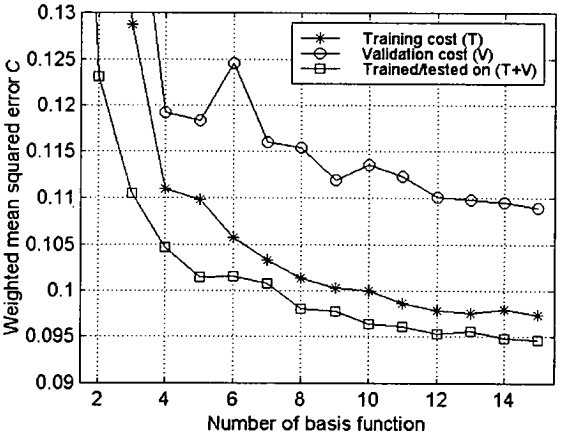


Figure 5.10: McHenry's method and MARS, costs on each dataset

(a) McHenry's method



(b) MARS



5.6.8 Discussion

While exhaustive selection is guaranteed to find the best set of parameters, its cost is prohibitive with the large number of parameters used here. McHenry's method appears to provide an alternative that is nearly optimal but allows sets of much larger size to be evaluated in only a few minutes' computation. However, omitting the second, more complex, check from McHenry's method makes it little better than cheaper methods. Forward selection, and variants such as forward-backward selection and multiple forward-backward selection, are quick but most do not find sets that are as good as McHenry's method; however, the datasets that they find, by being

less highly optimised to the training data, sometimes give a model that generalises better to unknown data.

The partitioning of the available data into separate training and test datasets is most useful for model order identification – finding the best size of parameter subset. For the linear fit using McHenry's method, order 2 seems to give the best cost and generalisation. Cross-validation illustrates that the non-linear MARS method appears to over-train for both low and high orders. MARS gives the lowest cost of all of the methods considered in this chapter when trained and tested on the combined dataset, but this tendency to over-train meant that the linear method was preferred.

5.7 Joint optimisation of parameter set and experiment fits

Sections 5.4–5.6 have introduced a number of methods for parameter selection and regression and applied these to estimation of raw and normalised MOS. This will not in general result in the optimum model when evaluated using the per experiment mapping technique described in section 5.3. This section considers the selection of a parameter set and fit that are jointly optimised for maximum performance, measured using the subjective test mapping, over the whole subjective test dataset (T+V).

5.7.1 Regression method

It is straightforward to extend the non-linear mapping process of section 5.3.4 to jointly optimise both the regression coefficients **a** (5-4) and mapping coefficients $\mathbf{b}_s$ (5-5). This is achieved by constructing a vector of the coefficients (5-29):

$$\mathbf{a}_c = \left[\mathbf{a}^T \quad \mathbf{b}_1^T \quad \cdots \quad \mathbf{b}_S^T\right]^T \tag{5-29}$$

Initialisation of the regression coefficients **a** and the mapping coefficients $\mathbf{b}_s$ for each subjective test s is performed separately, using normalised MOS to initialise **a**, then mapping the model output derived using this to each subjective test. The overall model (5-29) is then optimised using multi-dimensional simplex gradient descent with the cost function given in (5-8). To reduce redundancy, the constant offset coefficient in **a** is removed. This still typically leaves one degree of redundancy in the form of a scaling on y_0 which may be matched by scaling in each per experiment mapping; this is avoided by fixing the bounds within which $f()$ (5-4) must be monotonic, imposing an indirect cost constraint via the freedom of the output polynomials which are required to be monotonic in this range. This is sufficient to ensure that the optimisation produces an OSQ of similar magnitude to the initialisation values.

The improvement in cost obtained by gradient descent on the joint set of parameters described by (5-29) was found to be negligible: optimisation using 10,000 iterations of the

simplex algorithm reduced C by less the 0.1%, and was very slow. Conjugate gradient descent with line minimisation similarly was unable to make any useful improvement in cost.

This is thought to be because the minimisation of each $\mathbf{b}_s$ with respect to cost means that the gradient on $\mathbf{a}$ is zero. The small improvement observed was because the gradient descent methods that were used to compute each $\mathbf{b}_s$ were not absolutely optimal, as the simplex search terminated when the gradient fell below a small threshold value.

The process to evaluate the joint cost of a candidate parameter set and experiment mappings was obtained by using regression against normalised MOS to initialise $\mathbf{a}$, then computing each $\mathbf{b}_s$ using the monotonic polynomial method. This allows a number of cost criteria to be applied per experiment, such as a minimum value for correlation coefficient on a given set of tests.

For this final model training, the cost criterion that was used for parameter selection was maximum worst-case correlation coefficient. It is common for standards bodies to place requirements on worst-case performance: PESQ was required to exceed thresholds that ranged from 80–95%, depending on the design and structure of each subjective test. Furthermore, the worst-case performance is considered by many end-users to give a strong indication of the model's wider applicability.

5.7.2 Results

The results of performing model selection using linear regression with McHenry's selection algorithm are shown in Table 5.9. The cost criterion used in selection was $1-\varrho_{min}$ and the model was trained and tested on the combined dataset. Table 5.9 shows, for each model order, whether the resulting fit was monotonic, the weighted cost C according to (5-8), and the mean and worst-case correlation coefficient. For comparison, when selection was performed to optimise C, for M=2, the best ϱ_{mean} and ϱ_{min} were 0.9433 and 0.8599 respectively.

Table 5.9: Results of joint optimisation on best worst-case correlation

Order M	Monotonic	C (T+V)	ϱ_{mean} (T+V)	ϱ_{min} (T+V)
1	yes	0.0994	0.9317	0.7980
2	yes	0.0738	0.9503	0.8926
3	no	0.0732	0.9509	0.8957
4	no	0.0740	0.9501	0.8964

5.7.3 Final choice of model

In the previous section, Figure 5.9 and Figure 5.10 both indicate that the 2-parameter model gives best generalisation for this data. In Table 5.9 there is a small increase in ϱ_{min} for orders 3

and 4, but these models are no longer monotonic, and the figures for C and ρ_{mean} do not show a consistent improvement over the 2-parameter fit. Because of this, the final model chosen was the 2-parameter fit. Results from this model are presented in the next section. The distortion parameters used in this cognitive model are:

- P.AAM disturbance total, whole signal, p=3.5, 3.0, 2.0
- P.AAM asymmetric error, whole signal, p=1.0, 6.0, 2.0.

5.8 Results

This section presents results for the final model selected in the previous section, compared with PSQM [ITU-T P.861], PAMS 3.1 y_{LQ}, and PESQ [ITU-T P.862]. The models are evaluated against the full database of subjective tests described in Appendix D, which contains 2,119 conditions.

Figure 5.11 presents scatter plots comparing model quality score against MOS for each of the four models. This shows the shape of the relationship, and allows the distribution of errors to be visualised. Table 5.10 summarises the performance of the models in terms of mean and worst-case correlation coefficient, and RMSE in units of MOS, calculated per condition, after monotonic cubic regression to map objective score onto MOS for each subjective test. The estimated 95% confidence interval on the mean correlation coefficient, calculated from the standard deviation of the set of correlation coefficients for each model, is also shown in Table 5.10. Finally, Table 5.11 shows the distribution of absolute residual errors for the four models, again evaluated after monotonic cubic regression for each subjective test.

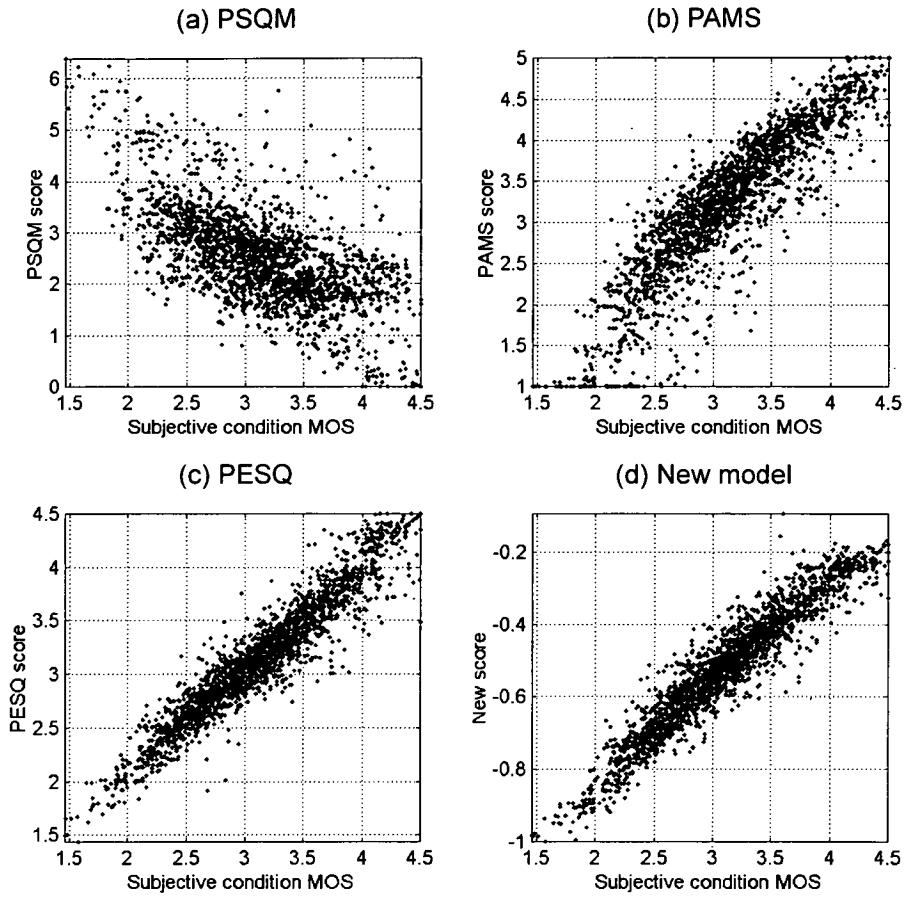
Table 5.10: Performance comparison of models

Model	Mean correlation and 95% confidence interval	Worst-case correlation	RMSE
PSQM	0.7780 ±0.0498	0.2796	0.381
PAMS	0.9295 ±0.0159	0.7642	0.226
PESQ	0.9435 ±0.0122	0.8101	0.199
New	0.9503 ±0.0087	0.8926	0.191

Table 5.11: Absolute residual error distribution after experiment mapping

Error magnitude		0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
Proportion of errors within range	PSQM (%)	44.41	73.34	87.49	94.81	98.07	99.34	99.91	99.95
	PAMS (%)	71.26	92.17	97.78	99.43	99.91	100.00	100.00	100.00
	PESQ (%)	74.14	93.77	99.10	99.81	100.00	100.00	100.00	100.00
	New (%)	75.70	94.67	99.34	99.86	99.95	100.00	100.00	100.00

Figure 5.11: Scatter plot of perceptual models against MOS



On almost all of the performance measures in Table 5.10 and Table 5.11, the new model performs best. There is one exception: a single outlier at more than 1.0 MOS away from the regression line. This can be seen in Figure 5.11(d) at (3.6, –0.1). This one outlier could be simply be reduced in magnitude without affecting the other conditions by placing an upper threshold on the model's quality score.

The results presented above can also be used to evaluate the overall effect of the innovations described in this thesis on model accuracy. Before this work began, PSQM [ITU-T P.861] was the state-of-the-art. All three of the models developed during this study, PAMS, PESQ and the New model, have mean performance that is significantly higher than PSQM, based on the means and confidence intervals shown in Table 5.10.

To compare PAMS, PESQ and the New model, a more powerful test must be used as the confidence intervals overlap slightly. A paired two-sided t-test [Duckworth 1968] was therefore applied to the sets of correlation coefficients to compare models. This showed the following:

- PAMS 3.1 is significantly more accurate than PSQM, with $P(T<t)=2.5 \cdot 10^{-7}$
- PESQ is significantly more accurate than PAMS 3.1, with $P(T<t)=0.005$

- The New model is very likely to be more accurate than PESQ, with P(T<t)=0.066 in the two sided test, or P(T<t)=0.033 in a one-sided test; however, the two-sided result would not be considered to show a significant difference between the models at the 95% confidence level.

5.9 Conclusions

This chapter has considered the use of a number of different regression, parameter selection and normalisation methods for training cognitive models.

The relationship between objective speech quality and MOS is in general non-linear and varies from test to test. The logistic function has been used to provide a linearising mapping for performance assessment. A new method of monotonic polynomial fitting was applied for this purpose. This preserves order and can curve more freely than the logistic function, allowing it to adapt to a wider range of curve shapes. As the polynomial cannot reach the same flatness as the logistic, it provides a more critical assessment of models' prediction ability.

Direct regression against MOS for the large set of subjective tests listed in Appendix D gives models that generalise badly, a consequence of systematic variations between tests. Non-linear normalisation of MOS per experiment provides a way to reduce this problem, and was found to produce a fit for a given parameter set that is very close to the optimum achievable when measured using the per experiment mapping process.

During model development it is convenient to generate large numbers of parameters and to select a small optimum set for use in the final mapping. Exhaustive search for parameter selection is only computationally tractable for small numbers of parameters, while forward and forward-backward selection were found to be fast but non-optimal. McHenry's parameter selection method was found to provide a good compromise between search time and closeness to the optimum performance.

For the dataset used in this study, it was found that non-monotonic non-linear methods have little benefit compared to linear regression. The Volterra, MARS and standard MLP methods were all found to be susceptible to over-training. The monotonic methods considered – parameter linearisation/bounding, the constrained MLP, and monotonic multivariate polynomial regression – only generalised slightly better. However, the last two of these are much slower to evaluate than the linear method, making them unsuitable for use in parameter selection, and the evidence from this data suggests that a 2-parameter linear fit gave the best generalisation.

Joint optimisation of the parameter set and per experiment mapping was performed over the whole dataset for the chosen order, and found to give a model with good average and worst-case correlation.

The results presented here indicate that the new P.AAM perceptual model has better performance than PESQ, particularly in the worst-case. This model exceeds 89% correlation with MOS for all 45 subjective tests. The residual error distribution indicates that there remain just over 5% of conditions for which the model is more than 0.5 MOS in error, and these should be considered in further work.

CHAPTER 6

Conclusions and further work

6.1 Conclusions

This thesis demonstrates that it is possible for perceptual models to accurately predict speech quality for the wide range of conditions outlined in the scope. Previous models, such as BSD, PSQM and early versions of PAMS, were limited to the assessment of speech coders, but the models developed during this study give high correlation with subjective MOS for a large database of subjective tests representative of most current telecommunications network technologies. This success has been due to the innovations in time-delay estimation, frequency response equalisation, and cognitive modelling, that were described in the previous chapters. These new methods were used in the development of PESQ, which replaced PSQM as the ITU standard speech quality measure. PESQ is now in widespread used in the telecommunications industry.

One of the main benefits of perceptual models is that they can be used to assess the perceived quality of non-linear, time-varying systems, where traditional signal processing metrics are of limited use. However, these properties also make it difficult to apply conventional techniques for time-delay and transfer function estimation, in particular where the end-to-end audio connection has poor phase stability due to clock jitter, low bit-rate coding, or channel errors.

The method of comparison of auditory transforms that is used in the perceptual models described in this thesis requires the delay to be known. To avoid slow iterative application of the model at all possible delays, methods to compute the delay directly are desirable. It was shown that cross-correlation-based techniques give inaccurate delay estimates in some critical cases. A new method based on constructing a smoothed, weighted histogram of short-term delay estimates was presented, and the results indicate that this leads to more accurate quality predictions.

Several network technologies, particularly VoIP, may cause changes in the end-to-end audio delay during a call. This seriously reduces the accuracy of perceptual models that assume constant delay. The dynamic time-warping method used by other authors was found to be of

little benefit for this application. The new method presented here to identify the delay for each speech utterance separately allows the model to take account of delay variations during silent periods, and gives a large improvement in performance for VoIP tests. A further new procedure described in this thesis, splitting utterances using the histogram method with a maximum-likelihood approach, is shown to provide a significant additional increase in model accuracy for variable-delay tests. Processing over delay changes to detect transients, and Hekstra's bad frame realignment process, both give small further improvements. Using these techniques, perceptual models can produce accurate quality scores for VoIP and other networks that may be subject to delay variation.

Linear filtering may be encountered in many elements of telephone networks, such as handsets or 2-wire analogue links, and usually introduces little perceived degradation. Because they consider linear and non-linear distortions in the same way, PSQM and other previous speech quality assessment models give very inaccurate results in the presence of filtering. It was shown that equalisation of the reference signal to the degraded signal can be used to address this problem. Because of the potential time-variance of systems in the scope of this thesis, linear methods may be highly biased and cannot be applied directly. The use of spectral difference and a new phaseless cross-spectrum method in the perceptual domain was described, and it was shown that the phaseless cross-spectrum method provides better noise rejection. The results indicate that estimating the frequency response of the system, and partially equalising the reference signal to the degraded signal, greatly enhances model accuracy for conditions that include filtering.

The non-linear variation of MOS between subjective tests poses an interesting challenge for training the output stage, or cognitive model, and for evaluating model performance. A new method for monotonic polynomial regression was described; this allows greater flexibility in the comparison between objective and subjective scores than the logistic method considered by most previous authors. The extension of this method to the normalisation of MOS was shown to improve prediction accuracy for the standard function approximation algorithms considered.

Over-training was found to be a significant risk, particularly for non-linear methods with many free coefficients. Imposing a monotonic constraint to preserve knowledge of the sign of the relationship between any distortion parameter and predicted quality can improve models' ability to generalise. A new weight processing method was presented that allows this to be enforced in the MLP.

A number of parameter selection methods were considered. Of these, McHenry's method was found to provide a good balance between accuracy and search size. This was applied in conjunction with linear regression to produce a new cognitive model for P.AAM, that was found to give higher overall and worst-case accuracy than PESQ.

6.2 Further work

It was shown in chapter 3 that Bayes' theorem can be applied to the identification of constant time-delay and of delay changepoints. This potentially is more powerful than the ad hoc procedures implemented for this study, and could enhance perceptual models' accuracy for variable delay conditions. Analysis of the performance of these methods in terms of delay and changepoint estimation error would also merit further study and may highlight further improvements.

While it was found that partial transfer function equalisation allows perceptual models to give more accurate scores in the presence of linear filtering, the majority of the data considered has been from typical network operating conditions, and very few pathological cases of filtering have been tested. This should be investigated further before these perceptual models are applied to testing terminal devices, rather than the current focus on testing networks that may have a non-flat frequency response.

This thesis has focused on monaural listening through a telephone receiver. Preliminary studies by the author indicate that it may be possible to apply the same auditory and cognitive model, with necessary changes to the input filtering, to the evaluation of quality using HATS as an acoustic receiver, or for wideband (16kHz sampling rate) telephony. In both of these cases, material is presented to subjects binaurally, over wideband headphones.

For the acoustic evaluation of hands-free telephones and/or noise at the listener using this method, binaural effects may be important, and this may necessitate substantial extensions to the cognitive model to take into account humans' remarkable ability to separate competing sounds and deconvolve the acoustic environment.

The model developed in the previous chapter shows higher correlation with MOS than its predecessors, but in some cases correlation is still only about 90%, and there are outliers where the model is more than 0.5 MOS in error. These cases need to be understood so that they can be addressed in further model development, or avoided when the model is used in the field.

Perceptual modelling for quality measurement is still relatively new, and many related problems remain unsolved. In audio testing, PEAQ is now five years old; applying the methods described in this thesis could help to produce an improved audio quality model. In telephony, non-intrusive measurement of speech quality is receiving strong commercial interest. Other problems that are important for current and next-generation telephony services include the objective measurement of conversation quality, particularly with large delay, and assessment of both streaming and conversational video.

Appendix A. List of publications and patents

The following tables list publications and patents that were produced as part of this study. The author's contribution to each is noted using the following key:

1. Main author responsible for more than 90% of the content.
2. Lead author responsible for at least 50% of the content.
3. Co-author responsible for less than 50% of the content.

The ICASSP 2000 paper (number 14; [Rix 2000b]), and the AES 109th paper (number 15; [Rix 2000c]) are reproduced in the following two appendices.

Num.	Publication	Note
1	Rix, A. W. and Hollier, M. P. *Robust design methodology for telephony assessment models.* ITU-T delayed contribution COM12-D031, February 1998.	1
2	Rix, A. W., Hollier, M. P. and Gray, P. *Predicting speech quality of telecommunications systems in a quality differentiated market.* 6th IEE Conference in Telecomms (ICT'98), Edinburgh. IEE Conference Publication 451, 156-160, March 1998.	2
3	Rix, A. W. and Hunt, T. J. *PAMS Trial Evaluation.* BT Systems Engineering document 421b02/T/007, March 1998.	1
4	Rix, A. W., Beamond, E. J., Hollier, M. P. and Gray, P. *Performance metrics for objective quality assessment systems in telephony.* ITU-T delayed contribution COM12-D079, November 1998.	2
5	Rix, A. W. and Hollier, M. P. *Comparison of speech quality assessment algorithms: BT PAMS, PSQM, PSQM+ and MNB.* ITU-T delayed contribution COM12-D080, November 1998.	1
6	Rix, A. W., Bourret, A. and Hollier, M. P. Modelling human perception. *BT Technology Journal,* 17 (1), 24-34, January 1999.	2
7	Rix, A. W., Reynolds, R. J. B. and Hollier, M. P. *Perceptual measurement of end-to-end speech quality over audio and packet-based networks.* 106th AES Convention, Munich, preprint 4873, May 1999.	2
8	Rix, A. W. *Comparison of opinion scales for subjective listening tests.* ITU-T delayed contribution COM12-D102, September 1999.	1
9	Rix, A. W. and Hollier, M. P. *Perceptual speech quality assessment from narrowband telephony to wideband audio.* 107th AES Convention, New York, preprint 5018, September 1999.	1
10	Rix, A. W., Reynolds, R. J and Hollier, M. P. *Robust end to end perceptual quality assessment of audio communication over packet-based networks.* IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'99), New Paltz NY, pp 39-42, October 1999.	1
11	Rix, A. W. *Advances in objective quality assessment of speech over analogue and packet-based networks.* IEE Data Compression colloquium, 99/150, November 1999.	1

Num.	Publication	Note
12	Rix, A. W., Hekstra, A. P., Beerends, J. G. and Hollier, M. P. *Performance of the integrated KPN/BT objective speech quality assessment model.* ITU-T delayed contribution COM12-D136, April 2000.	1
13	Beerends, J. G., Rix, A. W., Hekstra, A. P. and Hollier, M. P. *Proposed draft recommendation P.86x: Perceptual Evaluation of Speech Quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs.* ITU-T delayed contribution COM12-D140, April 2000.	2
14	Rix, A. W. and Hollier, M. P. *The perceptual analysis measurement system for robust end-to-end speech quality assessment.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Istanbul, (3), 1515-1518, June 2000.	1
15	Rix, A. W., Beerends, J. G., Hollier, M. P. and Hekstra, A. P. *PESQ - the new ITU standard for end-to-end speech quality assessment.* 109th AES Convention, Los Angeles, preprint 5260, September 2000.	2
16	Rix, A. W. *Results of quality assessment of wideband speech using PAMS.* ITU-T delayed contribution COM12-D001, January 2001.	1
17	Rix, A. W., Beerends, J. G., Hekstra, A. P. and Hollier, M. P. *Proposed modification to draft P.862 to allow PESQ to be used for quality assessment of wideband speech.* ITU-T delayed contribution COM12-D007, February 2001.	1
18	Reynolds, R. J. B. and Rix, A. W. Quality VoIP - an engineering challenge. *BT Technology Journal.* 19 (2), 23-32, April 2001.	3
19	Rix, A. W., Beerends, J. G., Hollier, M. P. and Hekstra, A. P. *Perceptual Evaluation of Speech Quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Salt Lake City, (2), 749-752, May 2001.	1
20	Reynolds, R. J. B. and Rix, A. W. Achieving VoIP voice quality. In Swale, R. (ed), *Voice over IP: systems and solutions,* 29-49. IEE, December 2001.	3
21	Rix, A. W. *P.AAM standardisation process.* ITU-T delayed contribution COM12-D083, May 2002.	1
22	Rix, A. W. *A new PESQ-LQ scale to assist comparison between P.862 PESQ score and subjective MOS.* ITU-T delayed contribution COM12-D086, May 2002.	1
23	Rix, A. W., Hollier, M. P., Hekstra, A. P. and Beerends, J. G. Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part I - Time-delay compensation. *Journal of the Audio Engineering Society,* 50 (10), 755-764, October 2002.	1
24	Beerends, J. G., Hekstra, A. P., Rix, A. W. and Hollier, M. P. Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part II - psychoacoustic model. *Journal of the Audio Engineering Society,* 50 (10), 765-778, October 2002.	3
25	Rix, A. W. *Analysis of P.862 PESQ scores using PESQ-LQ and alternative logistic mapping.* ITU-T delayed contribution COM12-D124, January 2003.	1
26	Beerends, J. G., Berger, J. and Rix, A. W. *Preliminary Results for the P.AAM benchmark models.* ITU-T delayed contribution COM12-D109, January 2003.	3
27	Rix, A. W., Berger, J. and Beerends, J.G. *Perceptual quality assessment of telecommunications systems including terminals.* AES 114th Convention, Amsterdam, preprint 5724, March 2003.	2

Num.	Publication	Note
28	Rix, A. W. *Comparison between subjective listening quality and P.862 PESQ score.* Workshop on Measurement of Speech and Audio Quality in Networks (MESAQIN'03), May 2003. http://wireless.feld.cvut.cz/mesaqin	1

Num.	Patent	Note
1	Reynolds, R. J. B., Hollier, M. P. and Rix, A. W. *Conversational testing of telecommunications equipment.* International patent application WO9853589, May 1997.	3
2	Rix, A. W., Reynolds, R. J. B., Hollier, M. P. and Sheppard, P. J. *Dynamic conversational testing of telecommunications equipment.* International patent application WO9853590, May 1997.	2
3	Rix, A. W., Reynolds, R. J. B., Hollier, M. P., Sheppard, P. J. and Beamond, E. J. *Measurement of speech signal quality for networks exhibiting variable delay.* International patent application WO0022803, October 1998.	2
4	Rix, A. W. *Neural network training process.* European patent application EP1065601, July 1999.	1
5	Reynolds, R. J. B., Rix, A. W., Hollier, M. P. and Gray, P. *Testing computer telephony.* International patent application WO0111854, August 1999.	3
6	Hollier, M. P., Gray, P., Reynolds, R. J. B. and Rix, A. W. *Testing transmission quality during silent intervals.* International patent application WO0193470, May 2000.	3
7	Gray, P., Reynolds, R. J. B., Rix, A. W. and Hollier, M.P. *In-service measurement of perceived speech quality by measuring objective error parameters.* International patent application WO0197414, June 2000.	3
8	Reynolds, R. J. B., Gray, P., Hollier, M. P. and Rix, A. W. *Method to reduce the distortion in a voice transmission over data networks.* International patent application WO0201824, June 2000.	3
9	Reynolds, R. J. B., Gray, P., Hollier, M. P. and Rix, A. W. *Method to assess the quality of a voice communication over packet networks.* International patent application WO0203633, June 2000.	3
10	Hollier, M. P., Gray, P., Reynolds, R. J. B. and Rix, A. W. *Optimum routing based on quality measurements.* European patent application EP1265446, June 2001.	3

Appendix B. ICASSP 2000 paper: *The Perceptual Analysis Measurement System for robust end-to-end speech quality assessment*

This paper was written as a summary of all of the main developments described in this thesis, applied to PAMS, and was presented in June 2000. ICASSP is one of the leading peer-reviewed conferences on general speech and signal processing.

4 pages

THE PERCEPTUAL ANALYSIS MEASUREMENT SYSTEM FOR ROBUST END-TO-END SPEECH QUALITY ASSESSMENT

Antony W. Rix and Michael P. Hollier
BT Advanced Communications Research
MLB 3/7 pp3, Adastral Park, Ipswich IP5 3RE, United Kingdom
antony.rix@bt.com

ABSTRACT

The perceptual analysis measurement system (PAMS) is an objective model designed to evaluate the perceived speech quality of telephone networks. Two key network properties, linear filtering and variable bulk delay, made previous models unsuitable for end-to-end measurement. This paper outlines the innovations that allow PAMS to take these properties into account. Variable delay occurs in packet-based transmission such as voice over IP. It is addressed by a time alignment algorithm that identifies delay changes in silent periods and during speech. Linear filtering, typically caused by analogue interfaces, is dealt with using a transfer function estimation and equalisation technique designed for robustness to non-linear distortions. A further development is the use of a constrained non-linear regression method to map error parameters to predicted subjective quality. Applications are discussed including the selection, optimisation and monitoring of telephone systems. Results are presented showing the performance of PAMS across 28 telephony subjective tests.

1. INTRODUCTION

Non-linear elements such as speech coders are becoming common in many audio communications systems. Conventional metrics such as frequency response or signal-to-noise ratio cannot reliably assess the quality of such systems. Synthetic test signals, such as sine waves or white noise, are inappropriate as coder behaviour is highly signal-dependent: for assessing speech coders, speech-like test signals are required. Traditionally, the only way to measure users' perception of quality was to conduct subjective tests [1–3], but these are expensive and unsuitable for applications such as commissioning or online monitoring.

Models were therefore developed to identify audible distortions through an objective process based on human perception. The basic concept of using perceptual masking in audio coding dates back to Schroeder et al. [4]. More recently, perceptual models for comparing an original and a degraded signal to assess quality were proposed, including those of Wang et al. [5], Beerends and Stemerdink [6, 7] and Hollier et al. [8, 9]. These led to ITU recommendations on two speech codec models, PSQM and MNB [10], and one audio quality model, PEAQ [11]. These objective models aim to predict the quality scores that would be given in a subjective test.

This paper focuses on the perceptual analysis measurement system (PAMS), based on the model proposed by Hollier [8, 9]. PAMS is designed for intrusive assessment of the speech quality of telephone networks, using speech or speech-like test signals. During development it became clear that codec assessment models such as [5, 7, 10] did not produce accurate scores when used in the field. Important end-to-end network behaviours that caused problems were linear filtering and bulk delay variations. More advanced error interpretation techniques were also needed to ensure accuracy across a wide range of distortion types.

The next two sections of this paper give an overview of PAMS, and describe some of its applications. The developments made to allow it to be used in end-to-end measurement are then outlined. Time alignment is discussed in section 4, focusing on variable bulk delay caused by packet-based transmission systems, and the methods used to address this are presented. Section 5 discusses linear filtering and the need for equalisation, summarising the algorithm used in PAMS. Results on the new model's performance are presented in section 6.

2. MODEL OVERVIEW

2.1 Structure of PAMS

The main elements of PAMS are shown in Figure 1. Before the signals can be compared they are aligned and equalised to account for bulk delay, variable delay, and linear filtering. Each signal is then filtered to simulate the response from the network junction, through a telephone handset, to the inner ear.

At the heart of PAMS (and all other perceptual quality assessment models) is the auditory transform. This maps the original and degraded signals to a time-frequency loudness representation, analogous to the transduction that occurs in the human auditory system. In this domain the difference between the two gives the error surface, a measure of audible errors caused by the system. This is interpreted by the perceptual layer,

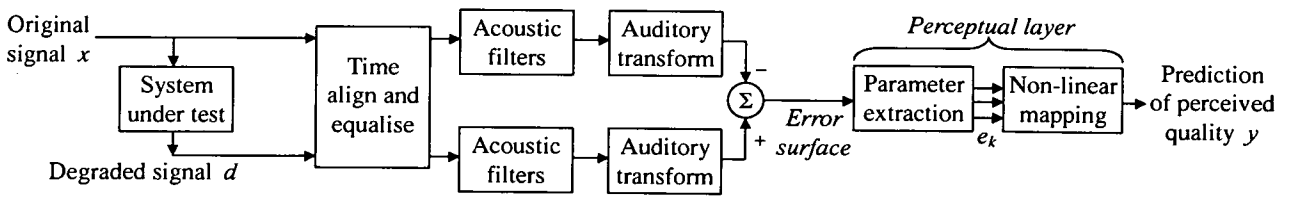


Figure 1: Structure of perceptual analysis measurement system.

which predicts the perceived quality of the system under test.

2.2 Auditory transform

The auditory transform used in PAMS is that of Hollier [8, 9], which was based on Wang et al. [5]. It consists of the following:

- 19-band Bark-spaced IIR perceptual filterbank;
- smoothing and downsampling;
- perceptual loudness mapping (phon and sone scale).

The use of a filterbank gives advantages in terms of temporal resolution, aliasing and dynamic range, compared to the FFT-based methods used in most other models [5–7, 10]. The output of the auditory transform is a spectrogram-like representation which approximates key large-scale psychoacoustic properties: temporal and simultaneous masking, and loudness scaling.

2.3 Robustness in the perceptual layer

The perceptual layer works in two stages [9]. The first extracts K error measures $e_1 \ldots e_K$, describing the amount and distribution of errors. These mimic perceptual processes, for example distinguishing between speech and silent periods, and between errors which add to or subtract from the original signal. The error measures are mapped to quality score y by a non-linear function calibrated to predict mean opinion score (MOS) using subjective test data. This paper considers the PAMS listening quality score, based on the listening quality absolute category rating (ACR) method [1]. Additionally, PAMS produces a score on the listening effort opinion scale using a similar mapping.

A model's robustness and generality depend critically on these mappings. Equation (1) shows the functional form used for PAMS. Individual error measures e_k are bounded then mapped by cubic polynomials $f_k()$. The whole fit is optimised by gradient descent with a cost-based constraint to ensure that each $f_k()$ is monotonically decreasing within bounds. This makes use of knowledge that as a given type of distortion gets worse, each e_k will by definition increase, and quality score y should fall. An optimal subset of K=10 error measures was chosen from a much larger candidate set using McHenry's selection method [12].

$$y = f_1(e_1) + \ldots f_K(e_K) + c, \quad \frac{df_k}{de_k} \le 0 \qquad (1)$$

The constrained mapping was compared with an unconstrained non-linear method, linear in the parameters (LITP) regression. This also used K error measures, their squares and cubes, with no product terms, giving $3K$+1 degrees of freedom. This was trained using the same optimal set of measures (K=10) and also an expanded set (K=20). Section 6 describes the training and test data sets and the performance metric (correlation coefficient after linearisation) used here. Table 1 gives the average and worst case results across the two data sets, for each method.

	K	*Training data mean/worst*	*Test data mean/worst*
Constrained PAMS method	10	0.939/0.867	0.963/0.924
LITP regression, 3rd order	10	0.935/0.901	0.953/0.920
LITP regression, 3rd order	20	0.948/0.929	0.953/0.893

Table 1: Comparison of results of regression methods

This shows that removing the monotonic constraint and adding more parameters both give improved correlation with the training data, but reduced correlation with the unseen test data set. The constrained method appears better able to generalise, as well as giving a more clearly understood input/output relationship.

3. APPLICATIONS

The following are examples of measurement applications for which PAMS has been used [13].

Equipment selection. Objective models can rapidly compare communications equipment or algorithms from different manufacturers. For example, PAMS was used to assess PSTN/IP gateways to evaluate their performance for a range of packet loss rates and in combination with other transmission equipment such as mobile networks. Subjective tests were then performed on a small subset of conditions to verify the conclusions.

Commissioning. Transmission systems often have unknowns such as choice of coder, bit-rate, level or buffer. PAMS allows the optimum combination to be found quickly, even if the differences are too small to be measurable in a subjective test. This allowed an Internet service provider to bring an IP telephony service to market in a very short time.

Monitoring. Together with a network of measurement devices to make test calls, PAMS has been used to monitor the quality of large and complex communications systems. This provides benchmarks of quality over time and as loading changes, and could identify problems before they affect customers.

4. VARIABLE BULK DELAY

4.1 Characteristics of packet-based speech transmission

To avoid conversational impairment in two-way communication, it is desirable to minimise end-to-end delay. However, packet switching often delays packets by different amounts, so packets have to be buffered to produce a continuous audio stream. Although this prevents signal loss it increases delay. Practical systems seek to balance these demands, tolerating occasional loss for lower delay, often with the aid of dynamic algorithms that resize buffers as packet arrival statistics change [12, 14].

This results in step changes to the bulk delay of the system. The most common delay changes appear to occur during silent periods to make buffer resizing imperceptible. However it is also possible for delay to change during speech; indeed, this can be a good way to deal with packet loss or late arrival.

In IP telephony trials, silent period delay changes as large as ±100ms were observed and changes of ±25ms were common [14]. However, in our tests, typical delay changes in silence that cause the algorithms of [10] to register a severe drop in score, equivalent to 1 MOS point, are 20ms for PSQM and 5ms for MNB [12]. As such delay changes are imperceptible, the models are significantly in error in this case, so we recommend avoiding these models if packet-based transmission could be encountered.

4.2 Time delay compensation in PAMS

To address the problem of changing bulk delay, a variable delay time alignment algorithm was developed. This is the first stage of PAMS and works to identify and account for delay changes. Buffer resizing is the most important effect and, as noted above, causes delay to change imperceptibly during silent periods. This

was modelled by identifying and aligning utterances separately [15]. Though simple, this proved to be highly effective and was implemented in version 2 of PAMS in August 1998 [13].

Dealing with delay changes during speech is more challenging. In the presence of coding and/or errors it is difficult to align signals, because the standard cross-correlation methods assume linear time invariance. A two-stage process was developed: crude alignment followed by fine resolution identification of the most likely delay. It was found that this was reliable even for quite short sections of speech, enabling an algorithm to locate delay variations within utterances to be produced [15].

Unlike silent period delay changes, delay variations during speech may be audible depending on where they occur. For example, deletion of part of a stationary unvoiced section is less perceptible than deleting a key transition or breaking the structure of a periodic voiced section. Identifying delay changes is not therefore enough: the model must also sample across the delay change to determine its subjectivity. This process was incorporated into version 3 of PAMS in September 1999.

The algorithm's sensitivity to delay changes in silence was tested by inserting or deleting silence between utterances in test files. As long as the surrounding speech was unaffected, PAMS scores remained constant for delay changes as large as ±500ms [14]. The delay statistics and changepoints are returned as diagnostic information to form part of an evaluation.

As variations in delay during speech are often audible, a model's sensitivity to them must be compared with subjective test data. The results presented in section 6 include several tests containing deletion or insertion causing delay changes during speech.

5. LINEAR FILTERING

Models for codec assessment [5, 7, 10] were intended to be used in the laboratory, with all-digital simulations or other highly controlled implementations. In such cases filtering is avoided, or if present it will be the same throughout a test. This meant that the models were able to neglect the effect of linear filtering.

However, analogue interfaces in telephony usually introduce filtering. Even if measurement devices are attached at digital points, networks may still contain analogue links. Transfer functions encountered in telephony often introduce bandlimiting, typically to the range 300–3,400 Hz. Additionally there is often some degree of filtering within the passband. For example, the send path of a handset from mouth to junction boosts high frequencies with a slope of about +10dB per decade, while a typical 2-wire connection shows a slope of –6dB per decade within the passband [14]. Temporal dispersion tends to be short, with impulse responses no longer than a few milliseconds.

Linear filtering of this magnitude has relatively little subjective effect compared to non-linear coding distortions. The early objective models made no distinction, with the result that conditions with filtering received much lower scores than unfiltered conditions, destroying the correlation between objective and subjective quality. For example, in one of the tests reported in section 6, PAMS has a correlation with subjective MOS of 0.895. If equalisation is disabled to make it behave like the earlier models, this correlation falls to 0.640.

In many cases linear and non-linear distortions cannot be observed separately; even when they can, it is undesirable to have to calibrate for filtering. So models need to be able to equalise the signals in the presence of potentially severe non-linear processing. The standard FFT-based transfer function estimation technique, based on cross-spectrum estimation, is unsuitable as it requires time invariance. It is also risky to compute the filter gain from the difference between the long-term spectra of input and output, as this is sensitive to noise. For PAMS version 3, a method was developed that performs identification/equalisation in the perceptual filterbank domain using a combination of cross-spectrum estimation and spectral difference. This is more robust than FFT-based methods because the filterbank envelope used is insensitive to short-term phase.

To illustrate this, Figure 2 shows transfer function estimates for an error-prone low bit-rate mobile condition. The solid line shows the estimate computed with a popular signal processing package using the conventional cross-spectrum technique, clearly unstable above 2kHz. This can be compared with the PAMS filterbank-based estimate (dash dot line), which is close, particularly in the passband, to the modified IRS send characteristic which is the actual filter present in this case [12].

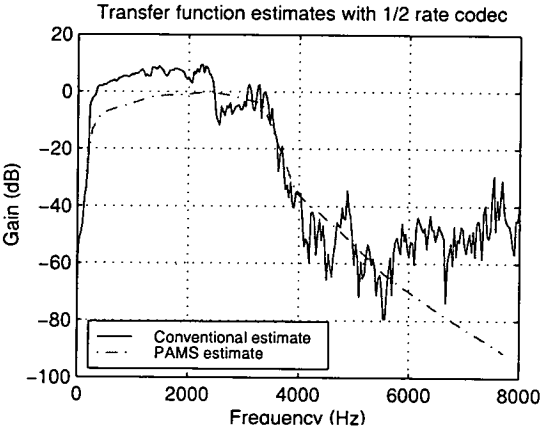


Figure 2: Transfer function estimation, coded channel.

A similar approach to transfer function identification and equalisation is taken in the PEAQ audio quality model [11].

6. PERFORMANCE RESULTS

The performance of PAMS as a predictor of subjective quality may be assessed by comparing its scores to subjective MOS. Many metrics can be used, such as RMS error, correlation coefficient or Spearman's rank correlation. Two sets of results are presented here: correlation coefficient (Figure 3) and distribution of residual errors (Figure 4). For both measures condition averaged scores are used, and a third-order monotonic polynomial is applied to linearise the relationship for each subjective test before the measures are calculated. This follows practice in the ITU-T for assessing objective models.

Figure 3(a) shows the 14 subjective tests that were used in the training of release 3.0 of PAMS. They are sub-divided into tests on mobile networks, fixed or international networks, voice over IP (VoIP), and multiple type tests; the last category covers tests that span at least two of the other categories. The data set is very

large, containing 8,480 files in 694 conditions. Four tests included background noise conditions. Ten were conducted in English and four in other European languages.

Figure 3(b) presents a further 14 subjective tests that were not used in the training process. These only covered the mobile and fixed/international categories, but contained even more data points: 13,252 files in 789 conditions. Two of the tests included background noise conditions. Eleven were conducted in English, one in French and two in Japanese. The average and worst case correlation coefficients for both the training and test data sets are given above on the first line of Table 1. All of these tests were subjectively scored using the listening quality ACR method [1].

Finally, Figure 4 plots the absolute residual errors for all 28 tests, tabulated at intervals of 0.25 MOS. This shows that for 95% of conditions the objective model (after mapping) was within 0.5 MOS of subjective opinion score. The RMS error is 0.236 MOS.

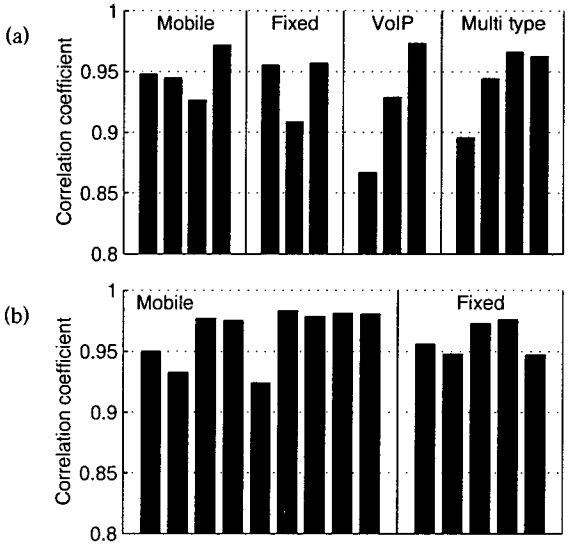


Figure 3: (a) PAMS performance for the 14 tests used in model calibration; (b) for 14 unseen subjective tests.

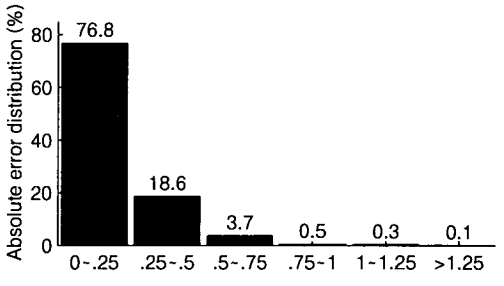


Figure 4: Distribution of residuals for all 28 tests.

7. CONCLUSION

Version 3 of PAMS improves on the performance of early codec assessment models and provides robustness for use in end-to-end measurement. It is able to deal with the key effects of filtering and variable delay. The perceptual layer functions well, providing good generalisation while retaining a monotonic relationship between error parameters and quality score. PAMS is being used successfully for speech quality assessment in many different applications.

8. ACKNOWLEDGEMENTS

The authors would like to thank all their colleagues who contributed to and reviewed this paper. Antony Rix is also supported by the Royal Commission for the Exhibition of 1851.

9. REFERENCES

[1] *Methods for subjective determination of transmission quality*. ITU-T Recommendation P.800, August 1996.

[2] *Subjective performance assessment of telephone-band and wideband digital codecs*. ITU-T Recommendation P.830, August 1996.

[3] *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*. ITU-R Recommendation BS.1116, July 1998.

[4] Schroeder, M. R., Atal, B. S. and Hall, J. L. "Optimizing digital speech coders by exploiting masking properties of the human ear". *Journal of the Acoustical Society of America*, 66 (6), 1647–1652, 1979.

[5] Wang, S., Sekey, A. and Gersho, A. "An objective measure for predicting subjective quality of speech coders". *IEEE Journal on Selected Areas in Communications*, 10 (5), 819–829, 1992.

[6] Beerends, J. G. and Stemerdink, J. A. "A perceptual audio quality measure based on a psychoacoustic sound representation". *Journal of the Audio Engineering Society*, 40 (12), 963–974, 1992.

[7] Beerends, J. G. and Stemerdink, J. A. "A perceptual speech-quality measure based on a psychoacoustic sound representation". *Journal of the Audio Engineering Society*, 42 (3), 115–123, 1994.

[8] Hollier, M. P., Hawksford, M. O. and Guard, D. R. "Characterisation of communications systems using a speech-like test stimulus", *Journal of the Audio Engineering Society*, 41 (12), 1008–1021, December 1993.

[9] Hollier, M. P., Hawksford, M. O. and Guard, D. R. "Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain". *IEE Proc. Vision, Image and Signal Processing*, 141 (3), 203–208, 1994.

[10] *Objective quality measurement of telephone-band (300–3400 Hz) speech codecs*. ITU-T Recommendation P.861, February 1998.

[11] *Method for objective measurements of perceived audio quality*. ITU-R Recommendation BS.1387, January 1999.

[12] Rix, A. W., Reynolds, R. and Hollier, M. P. "Robust perceptual assessment of end-to-end audio quality", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 39–42, October 1999.

[13] PAMS website: http://www.labs.bt.com/people/rixa/pams/

[14] Rix, A. W., Reynolds, R. and Hollier, M. P. "Perceptual measurement of end-to-end speech quality over audio and packet-based networks". *106th Audio Engineering Society Convention*, pre-print no. 4873, May 1999.

[15] Reynolds, R. and Rix, A. W. "Measurement of signal quality". *International Patent Application* PCT/GB99/03236.

Appendix C. AES 109th convention paper: *PESQ – the new ITU standard for end-to-end speech quality assessment*

The author co-wrote this paper with Beerends, Hekstra and Hollier to summarise the development, structure and performance of PESQ, and presented it to the Audio Engineering Society in September 2000. The paper was later extended and published in October 2002 in two parts [Rix 2002b, Beerends 2002] in the *Journal of the Audio Engineering Society*, which has been the main record on perceptual quality modelling.

18 pages

PESQ – the new ITU standard for end-to-end speech quality assessment

Antony W. Rix[1], John G. Beerends[2], Michael P. Hollier[1] and Andries P. Hekstra[2]

[1] *BT Advanced Communications Research, B54/86 Adastral Park, Ipswich IP5 3RE, United Kingdom*
[2] *Royal PTT Nederland NV, NL-2260 Leidschendam, The Netherlands*
e-mail: awr@iee.org

This paper describes a new model for perceptual evaluation of speech quality (PESQ). This model is based on an integration of the perceptual speech quality measure (PSQM99) and the perceptual analysis measurement system (PAMS). PESQ is currently a draft ITU-T recommendation P.862, and is expected to replace P.861. PESQ provides a new international standard for objective assessment of speech codecs and end-to-end measurement of telephone networks.

0. Introduction

Speech codecs and other non-linear elements are becoming common in many audio communications systems. Conventional signal processing measures, such as frequency response or signal-to-noise ratio, cannot be reliably used to assess the perceived quality of these complex, non-linear systems. Likewise common synthetic test signals, such as impulses, white noise or sine waves, are inappropriate for testing speech codecs as their behaviour is highly signal-dependent: instead, speech-like signals must be used.

Until recently the only way to measure users' perception of the quality of such systems was to conduct a subjective test [1–3]. However, subjective tests are expensive and slow, and cannot be used in certain applications such as in-service monitoring. Objective models, based on human perception, were therefore developed with the aim of predicting the results of subjective tests.

Perceptual masking in audio coding was first proposed by Schroeder et al. [4], who suggested a model to estimate the audibility of coding noise. Brandenburg extended this to give a measure of noise to masking ratio (NMR) [5]. The concept of comparing internal loudness representations to determine perceived errors was introduced by Karjalainen [6].

Beerends and Stemerdink published several models based on comparison of internal representations to give a single measure of quality for audio or speech codecs [7–11]. Their method for assessing narrowband speech codecs, the perceptual speech quality measure (PSQM) [10], was selected after a competition held by the International Telecommunication Union (ITU-T) and adopted as ITU-T recommendation P.861 [12]. Their method for audio codecs, the perceptual audio quality measure (PAQM) [8, 11] was combined with a number of other audio models to produce a new model known as perceptual evaluation of audio quality (PEAQ), which has recently become ITU-R recommendation BS.1387 [13–15].

Wang et al. also used the comparison of internal representations to give a single measure of speech quality, the Bark spectral distortion (BSD) [16]. It was refined by Hollier, incorporating a bank of linear filters for spectral analysis and by taking account not only of the amount, but also the distribution of audible distortions [17, 18]. This model became the psychoacoustic core of the perceptual analysis measurement system (PAMS) [19, 20], a model designed for assessing telephone networks as well as speech codecs.

Codec assessment models such as PSQM [12] had limitations which made them unreliable when used in certain applications, especially for systems that include linear filtering and/or delay variations. After an ITU-T competition to select a new end-to-end speech quality assessment model, the two algorithms with the highest overall performance in the competition, PAMS and PSQM99 (an updated version of PSQM),

were combined. The new model is known as perceptual evaluation of speech quality (PESQ). Following validation against a large number of subjective tests including real network measurements, PESQ was determined in May 2000 as a new draft ITU-T recommendation P.862 [21]. It is expected that P.862 will replace P.861 early in 2001.

Section 1 of this paper presents the background of subjective and objective quality assessment, introducing PSQM and PAMS and describing the development of models by the ITU-T. In section 2, some conditions in which previous models gave inaccurate results are described, including variable delay and filtering. Section 3 gives an overview of the structure of PESQ. Results comparing PESQ with P.861 are given in section 4. Section 5 describes the range of conditions for which information on the performance of PESQ is currently available and summarises the areas in which further work remains to be done, then conclusions are drawn in section 6.

1. Background and development

1.1 Subjective evaluation of speech quality

The methodology of subjective testing for speech quality is set out in ITU-T recommendations P.800 and P.830 [1, 2]. These describe methods for evaluating one-way listening quality as well as two-way conversational quality. Listening tests are the most common because they are cheaper and easier to conduct than conversational tests; however, listening-only testing cannot take account of factors which only affect conversation, specifically conversational level, sidetone, talker echo and round-trip delay. This paper considers objective models designed to predict the results of listening-only subjective tests.

Several different subjective rating methods are defined in [1]. The simplest is the absolute category rating method (ACR). In this, subjects hear a number of degraded recordings, and are prompted to vote on each one according to an opinion scale such as the 5-point listening quality (LQ) scale shown in Table 1. Degradation and comparison category rating (DCR, CCR) methods are also described in [1]. In the DCR method the subjects hear first an undistorted reference followed by the degraded recording, and then vote on their opinion of the audibility and annoyance of the degradation. In the CCR method the reference may be presented either before or after the degraded recording, and subjects are asked to give their opinion on how much better, or worse, is the second file compared to the first. The ACR method with the LQ opinion scale is the most commonly used method in telecommunications assessment, and was the primary focus of development of PESQ.

Quality of the speech	*Score*
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 1: Listening quality opinion scale [1].

There are a number of common factors in subjective listening tests for telecommunications.

- Listening is in a quiet room with a controlled noise level.
- Subjects listen through a telephone handset with a standard response.
- Recordings are typically 8s long and consist of a pair of unrelated sentences.
- Tests are performed with speech from several different talkers (typically two male, two female) for each coding condition.
- Subjects are non-expert.

Once the test is complete the votes are averaged across subjects to give a mean opinion score (MOS). The quality of each condition is given by the condition MOS. In some cases it may also be useful to compute statistics for each file (file MOS) or each talker (talker MOS) within a given condition.

1.2 Perceptual speech quality measure (PSQM)

PSQM was developed by Beerends and Stemerdink from ideas first presented in [7] as a model optimised for the assessment of speech codecs [10, 12]. An overview of the structure of PSQM is shown in Figure 1. The core of the model is an auditory transform that models key psychophysical processes in the human auditory system. This computes a spectrogram-like internal representation of loudness in time and frequency as follows [12]:

- Short-term Fourier transform (STFT) with 50% overlapping Hann windows, 32ms long.
- Frequency warping of short-term power spectrum to 56-band Bark scale.
- Local scaling – partial equalisation of degraded signal to reference based on power of each frame, to account for low-frequency gain modulation.
- Filter with handset receive characteristic.
- Add Hoth noise.
- Loudness warping to a compressed Sone loudness scale.
- Loudness scaling to equalise degraded signal to same total loudness as reference signal in each frame.

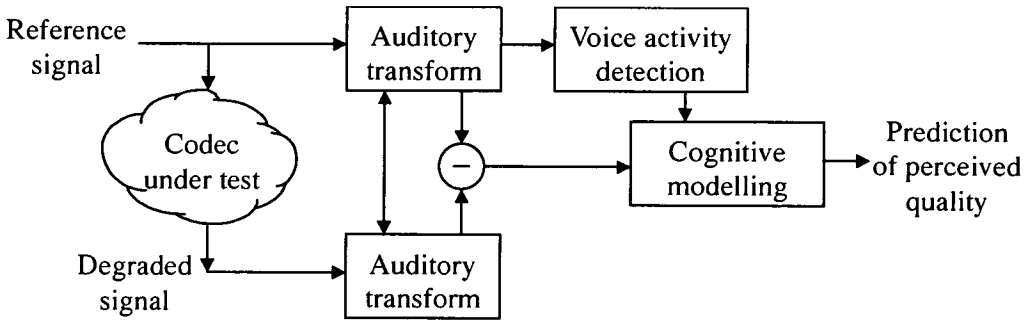


Figure 1: Structure of PSQM

Early approaches to estimating a single quality score were based on the average distance between the transforms of the reference and degraded signal [6, 7, 16]. PSQM introduced a "cognitive model" to interpret the difference between the transforms. This improves correlation between objective score and subjective MOS by modelling two effects: asymmetry and different weighting for speech and silence.

The asymmetry effect is caused by the fact that when a codec distorts the input signal it will in general be very difficult to introduce a new time-frequency component that integrates with the input signal, and the resulting output signal will thus be decomposed into two different percepts, the input signal and the distortion, leading to clearly audible distortion [22]. However, when the codec leaves out a time-frequency component the resulting output signal cannot be decomposed in the same way and the distortion is less objectionable. This effect is modelled in PSQM by multiplying the disturbance by a correction factor using the power ratio between the output signal and the input signal at a certain time-frequency point as a measure of "newness" of this component.

The fact that disturbances that occur during speech active periods are more disturbing than those that occur during silent intervals is modelled by a weighting factor that can be adjusted to the context of the experiment. Details of this procedure can be found in [12].

In the 1993-1996 study period ITU-T study group 12 compared five different objective models, LPC Cepstral Distance, Information Index, Coherence Function and Expert Pattern Recognition and PSQM, on

their ability to predict subjective quality scores. PSQM showed the highest performance on distortions which had not been used during training, and it was accepted in 1996 as ITU-T recommendation P.861 for objective quality measurement of narrowband speech codecs [12]. P.861 includes a detailed definition of its scope, which does not include live network testing or conditions in which variable delay (time warping) can occur. It also describes how the reference and degraded signals must be aligned in time and equalised to a fixed active speech level, corresponding to the standard level used in subjective listening tests.

Further work took place between 1996 and 1999 to improve PSQM and make it suitable for end-to-end testing of real networks, leading to a new version of the model, PSQM99, that incorporates time and level alignment routines.

1.3 Perceptual analysis measurement system (PAMS)

PAMS is an enhanced version of the model proposed by Hollier [17, 18]. This differs from the model of Wang et al. [16] by the use of a bank of linear filters – as opposed to the STFT – to implement time-frequency transformation and by a process termed the perceptual layer to interpret the error surface. Further development took place to provide time alignment, level alignment and equalisation functions essential for use in end-to-end measurement, and the perceptual layer was extended [20].

The model begins with time alignment, using a multi-stage process to align the reference and degraded signals. The signals are divided into sections known as utterances. Delay changes – for example due to packet-based transmission such as IP telephony – are identified. The signals are both equalised to a standard reference listening level corresponding to 79 dB SPL [2]. The auditory transform is then performed as follows.

- Input filter to model the response of the telephone handset, ear coupling and ear canal.
- Bank of linear filters to transform the signal to 19 Bark-spaced perceptual frequency bands.
- Computation of smoothed power envelope for each frequency band in 4ms frames.
- Transfer function estimation and equalisation to account for linear filtering in the system under test; the reference signal is partially equalised to the degraded signal.
- Mapping to phon loudness scale.
- Mapping to a Sone loudness scale (both phon and Sone loudness values are used in the model).

A number of error parameters are computed based on the auditory transforms of the reference and degraded signals, giving a measure of the amount of different classes of distortion. These are averaged in time and are mapped to quality score through a non-linear function, preserving a monotonic relation between each parameter and quality score. This process is outlined in Figure 2. Two quality measures are computed, one on the ACR listening quality opinion scale (Table 1) and the other on the ACR listening effort opinion scale [1].

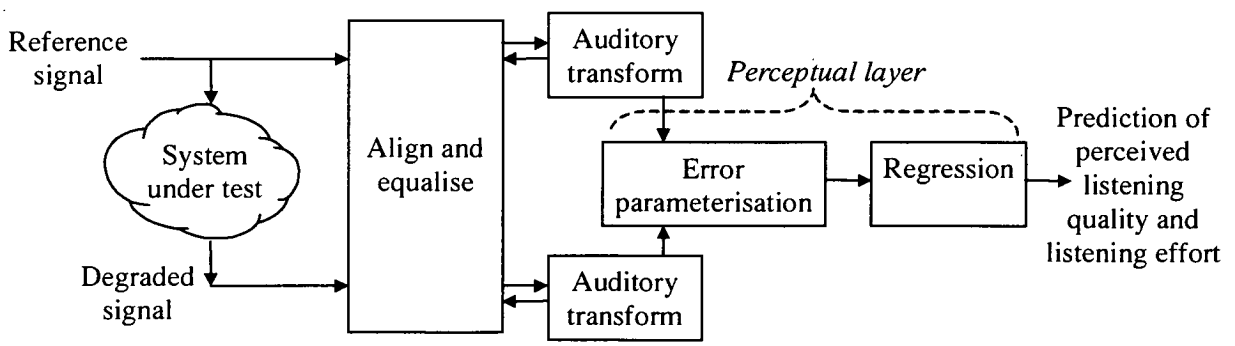


Figure 2: Structure of PAMS

The components that implement variable delay time alignment and transfer function equalisation were designed to enable PAMS to be used in end-to-end measurement applications [19, 20]. With appropriate changes to the input filter to model headphone listening – but with no alterations to the underlying auditory model and perceptual layer – PAMS has also been extended to testing monophonic wideband telephony at 16kHz sample rate [23].

1.4 Further development of perceptual models in the ITU-T

Following the identification of certain conditions in which P.861 PSQM was found to show poor correlation with subjective opinion, another algorithm, Measuring Normalizing Blocks (MNB), was proposed. This was added to P.861 as an informative – but not binding – appendix II in 1998 [12]. However, neither P.861 PSQM nor MNB were found to be suitable for end-to-end measurement of networks. Particular problems were observed with conditions that introduce significant linear filtering, variable delay, or additive background noise. To address these problems, a competition to select a new end-to-end assessment model was conducted by ITU-T study group 12 between September 1998 and March 2000. The competition was unable to identify a single outright winner as it was difficult to distinguish the two models with the highest overall performance. These models, PAMS and PSQM99, were combined and further work was carried out to meet a demanding set of requirements.

The new model, perceptual evaluation of speech quality (PESQ), was found to meet the requirements and to have significantly better performance than P.861 PSQM and MNB, even for tests including only speech codecs. PESQ was therefore determined in May 2000 by ITU-T study group 12 as a new draft ITU-T recommendation P.862 [21]. It is anticipated that P.862 will be approved early in 2001, at which point P.861 is expected to be withdrawn.

2. Weaknesses of previous models

2.1 Variable delay (time warping)

2.1.1 Processes causing delay variation

Large delays impair two-way conversation, so it is desirable to minimise end-to-end delay. However packet-based transmission often leads to each packet being delayed by a different amount. A buffer is required to iron out these delay variations and produce a continuous audio stream. It is necessary to balance the length of the buffer – a major addition to end-to-end delay – with the possibility of packet loss.

Two different processes appear to lead to delay variations in voice over IP (VoIP). Dynamic buffer resizing during silence is a common method for dealing with time-varying packet delay by changing the buffer length – and hence delay – during silent intervals. In a less frequent effect, large changes in packet delay can cause buffers to overrun or become empty, leading to delay changes during speech.

The magnitude of delay variations encountered in measurements of delay changes on two packet-based networks are presented in Figure 3. This shows the distribution of changes in end-to-end (audio) delay during a series of 16-second measurements on each type of network. Figure 3(a) gives results from a field trial of a PC-based VoIP system. Figure 3(b) shows the corresponding distribution measured across a PSTN to Internet Gateway system. Delay changes of up to 100ms in magnitude were encountered [19].

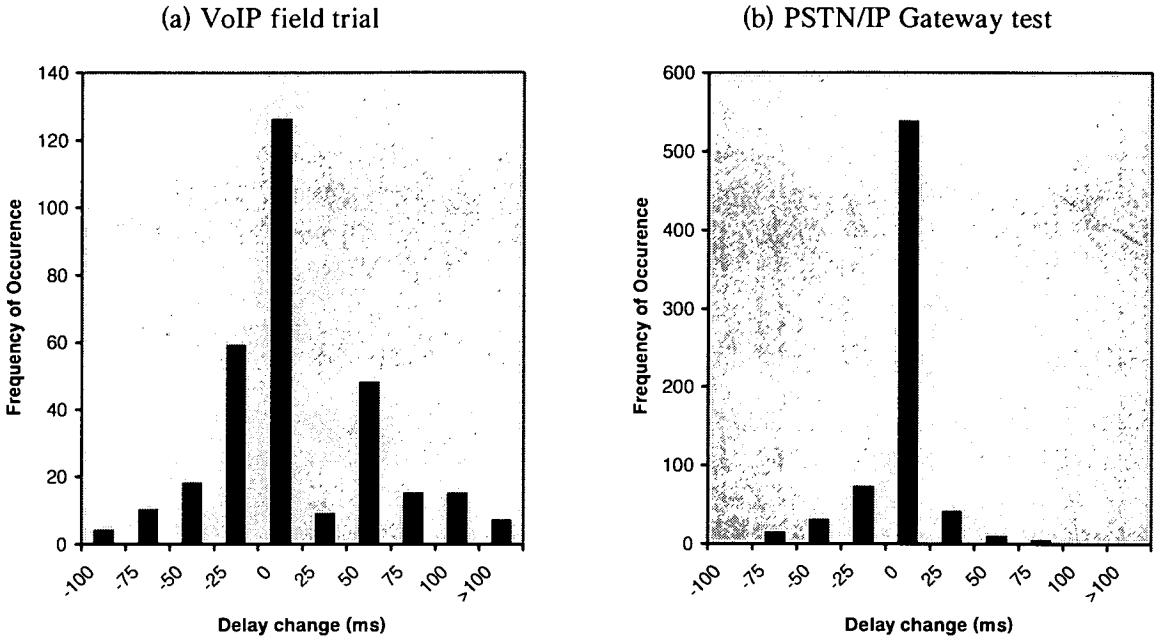


Figure 3: Delay variations measured in voice over IP network connections

2.1.2 Effect of variable delay on perceptual models

The sensitivity to variable delay of codec assessment models such as PSQM and MNB was evaluated in two ways. In the first investigation delay changes in silent periods were introduced. This models the effect of dynamic buffer resizing during silence. Delay changes such as these are not noticed by subjects unless they are very large (e.g. 0.5s) or cause insertion or deletion during speech events. However, ignoring delay variations causes an objective model to compare different sections of the reference and degraded signals. As speech signals are highly time-varying, this causes large false errors to be measured. It was found that a delay change of 20ms is sufficient to cause PSQM [12] to measure a drop in quality equivalent to approximately 1 MOS. MNB [12 appendix II] is even more sensitive, requiring a change of only 5ms to cause an equivalent drop in quality [19].

The second investigation was based on a subjective test in which the distortions included delay changes. The correlation between subjective and objective score was measured using the procedure described in section 4.1 and is summarised in Table 2. The models of P.861, PSQM and MNB, show very low correlation with subjective opinion and clearly give meaningless scores in this test. However, PESQ, which is designed to take variable delay into account, shows a much higher correlation.

Model	*Correlation*
PSQM [12]	0.260
MNB [12]	0.363
PESQ	0.932

Table 2: Correlation between subjective and objective score for VoIP variable delay test. Per condition, after monotonic 3rd-order polynomial mapping.

2.2 Linear filtering

2.2.1 Frequency response of typical network components

Many components used in telephony introduce significant amounts of linear filtering. This can occur in the acoustic path, at 2-wire line interfaces, or even in speech codecs. To give an idea of the magnitude of filtering that occurs in end-to-end measurement, Figure 4 shows the frequency response of two typical network components. Figure 4(a) shows the modified IRS send characteristic [2], which represents the frequency response from mouth to junction through a typical telephone handset. Figure 4(b) shows the response measured by a typical test device, with 2-wire interfaces, on a telephone connection in the UK [19]. These responses are quite typical and show that gain within the 300–3,400Hz passband can vary by ±10dB.

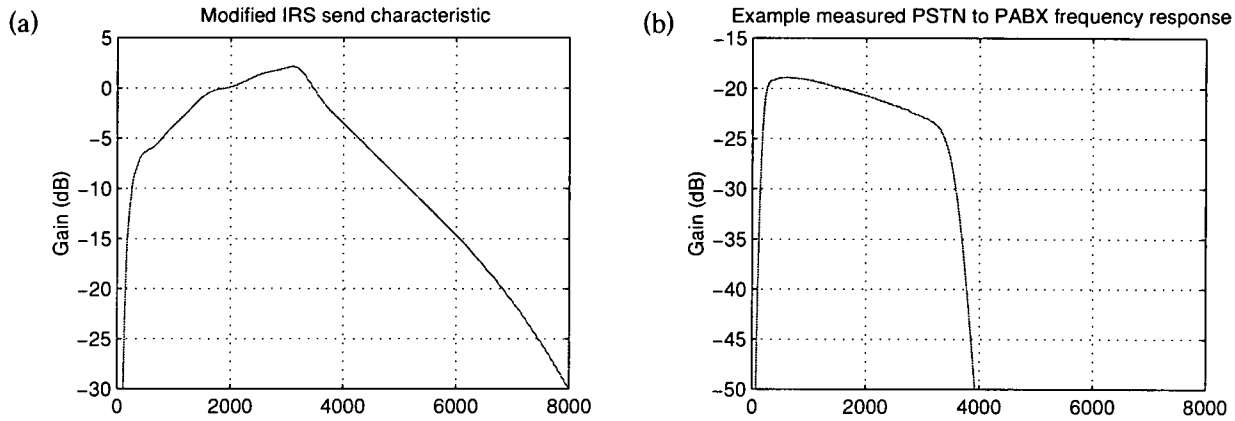


Figure 4: Frequency responses of telephone network components

2.2.2 Effect of filtering on perceptual models

Although linear filtering does have some subjective effect, it is generally much less significant than non-linear coding distortion. The early perceptual models such as BSD [16], PSQM and MNB [12] made no distinction and therefore measure large errors due to filtering alone. It is recognised that perceptual models for use with end-to-end audio systems must give less weight to linear distortions. This is usually achieved by equalising the reference signal to the degraded signal. Note that standard linear transfer function estimation/equalisation techniques cannot normally be applied because they are unstable with low bit-rate speech codecs [19]. Several approaches can ensure that audible filtering is not completely ignored. Partial compensation eliminates most of the effect but leaves some residual distortion to be measured by the perceptual model; this is the method used in PAMS, PSQM99 and PESQ. If the filtering is fully compensated, separate linear distortion measures can be used as part of the final regression to subjective MOS, the method used in PEAQ [13, 14].

The effect of filtering is illustrated by the correlation, given in Table 3, between subjective and objective MOS for a subjective test on low bit-rate mobile codecs. In half of the conditions in this test an IRS send filter similar to that of Figure 4(a) was applied; in the other conditions there was no filtering. PSQM and MNB [12] perform quite badly because the objective scores given to the filtered conditions are very different from the unfiltered conditions. PESQ take good account of the filtering, resulting in a much higher correlation than the models of P.861 [12].

Model	*Correlation*
PSQM [12]	0.616
MNB [12]	0.626
PESQ	0.914

Table 3: Correlation between subjective and objective score for mobile codec test with filtering. Per condition, after monotonic 3rd-order polynomial mapping.

2.3 Gain variation

Although uncommon in today's telephone networks, it is possible for speech to be subject to forms of low-frequency amplitude modulation. This can occur with automatic gain control (AGC), in which speech levels are dynamically adjusted towards a standard level. The intention is to cancel the effect of variable losses in subscriber equipment or level variations between different countries' networks. Sometimes, however, undesirable gain changes occur as a consequence of background noise or normal variations in vocal level. Most current AGC systems change level slowly, and only in silent periods, but some systems have been found to operate during speech.

The subjective effect of AGC is usually quite limited. Because speech is naturally time-varying, it is difficult to detect gain changes of less than 3dB, and gain changes of up to 10dB are not generally very disturbing as long as audible discontinuities are avoided. However, as perceptual models are based on a comparison of internal loudness representations, it is necessary to track and equalise gain variations otherwise large errors would be measured for a relatively inaudible effect.

Models differ widely in their approach to modulation. It is essentially ignored by MNB. PAMS identifies and cancels out gain changes only if they occur during silence, but gain changes during speech cause large errors to be measured. PSQM, PSQM99 and PESQ adaptively track envelope changes frame-by-frame, cancelling the effect with a lag to ensure that some errors are measured due to gain changes.

A number of tests used in the validation of PESQ contained gain variation; some of these results are reproduced in section 4.2. It appears that the adaptive method employed in PESQ takes good account of the subjective impact of gain variation.

2.4 Temporal clipping

A general term for the replacement of speech by silence, temporal clipping can occur with many different networks, including international, mobile and VoIP. The most common type is front-end clipping, where the first few milliseconds of speech are not transmitted. Back-end clipping is the corresponding effect where the end of a speech utterance is truncated. This is usually the result of voice activity detection errors with discontinuous transmission in silence (DTX), where transmission ceases when speech becomes inactive to free up capacity. DTX is usually accompanied by comfort noise insertion, which aims to re-create background noise of similar spectrum so that the listener does not notice the effect. Clipping may also occur during speech, for example if a packet is lost and is replaced by silence.

Different forms of temporal clipping appear to have very different subjective effect. Front-end clipping often has little perceived effect; in some tests, even 50ms of front-end clipping was not noticed by the subjects. However, in other tests where semantically important parts of speech were deleted by front-end clipping or by packet loss, as little as 10ms of clipping is found to be perceptually significant. It is clearly impossible for an objective model to predict these conflicting effects. Most current models predict that temporal clipping is significant, giving a drop of quality in the region of 0.5 MOS for around 50ms of front-end clipping, and this is probably the best that can be done with this type of distortion.

3. Structure of PESQ

An overview of PESQ is shown in Figure 5. The model begins by level aligning both signals to a standard listening level. They are filtered (using an FFT) with an input filter to model the telephone handset. The signals are aligned in time and are then processed through an auditory transform similar to that of PSQM. Part of the transformation involves equalising the signals for the frequency response of the system and for gain variation. The difference between the transforms of the reference and degraded signals is known as the disturbance. This is processed to extract two distortion parameters, which are aggregated in frequency and time and mapped to a prediction of subjective MOS. The details of the time alignment, transformation and disturbance processes are discussed in the following sections.

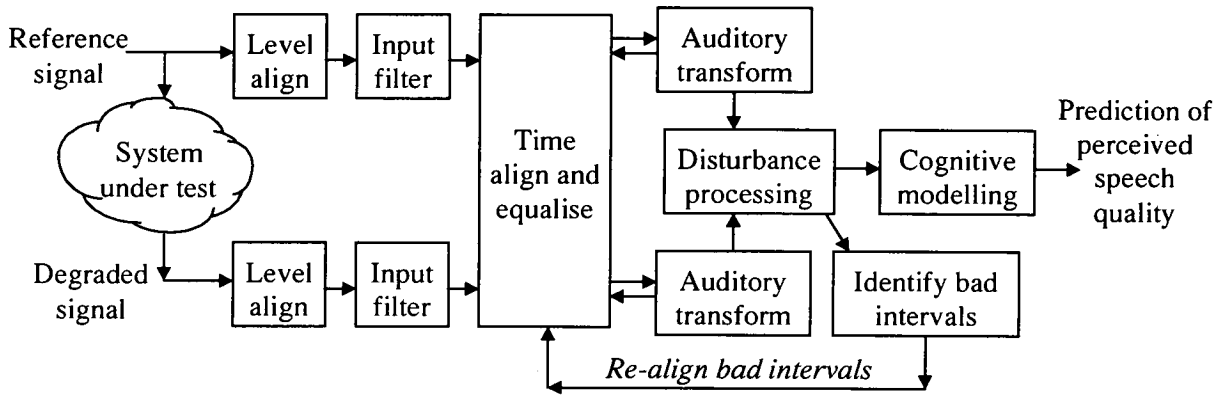


Figure 5: Structure of PESQ

3.1 Time alignment

The time alignment of PESQ assumes that the delay of the system is piecewise constant. Delay changes are allowed in silent periods (where they will normally be inaudible) and in speech (where they are usually audible). The signals are aligned using the following steps [20].

- Narrowband filter applied to both signals to emphasise perceptually important parts. These filtered signals are only used for time alignment.
- Envelope-based delay estimation.
- Division of reference signal into utterances.
- Envelope-based delay estimation for each utterance.
- Fine correlation histogram-based delay identification for each utterance.
- Utterance splitting and re-alignment to test for delay changes during speech.

The result is a number of sections with a given start, end, and delay. This is translated to a frame-by-frame delay for use in the auditory transform.

3.2 Auditory transform

The auditory transform of PESQ is a psychoacoustic model which maps the signals into a representation of perceived loudness in time and frequency.

Bark spectrum. A STFT with a Hamming window is used to calculate the instantaneous power spectrum in each frame, for 50% overlapping frames of 32ms duration. This is grouped without smearing into 42 bins, equally spaced in perceptual frequency on a modified Bark scale similar to that of PSQM [12].

Frequency equalisation. The mean Bark spectrum for active speech frames is calculated. The spectral difference gives an estimate of the transfer function, assuming that the system under test has a constant

frequency response. The reference is equalised to the degraded signal using this response, with bounds to limit the equalisation to ±20dB.

Equalisation of gain variation. The ratio between the audible power of the reference and the degraded in each frame is used to identify gain variations. This is filtered with a first-order low-pass filter, and bounded, then the degraded signal is equalised to the reference.

Loudness mapping. The Bark spectrum is mapped to (Sone) loudness, including a frequency-dependent threshold and exponent. This gives the perceived loudness in each time-frequency cell.

3.3 Disturbance processing and cognitive modelling

The absolute difference between the degraded and the reference signals gives a measure of audible error. In PESQ, this is processed through several steps before a non-linear average over time and frequency is calculated.

Deletion. If deletion occurs (a negative delay change) there will be a section which overlaps in the degraded signal. If the deletion is longer than half a frame, the overlapping sections are discarded.

Masking. Masking in each time-frequency cell is modelled using a simple threshold below which disturbances are inaudible; this is set to the lesser of the loudness of the reference and degraded signals, divided by four. The threshold is subtracted from the absolute loudness difference, and values less than zero are set to zero. Methods for applying masking over distances larger than one time-frequency cell were examined with earlier versions of PSQM and PSQM99, but did not improve overall performance [9], and were not used in PESQ.

Asymmetry. Unlike P.861, PESQ computes two different error averages, one without and one with an asymmetry factor. The PESQ asymmetry factor is calculated from a stabilised ratio of the Bark spectral density of the degraded to the reference signals in each time-frequency cell. This is raised to the power 1.2 and is bounded with an upper limit of 12.0. Values smaller than 3.0 are set to zero. The asymmetric weighted disturbance, obtained by multiplying by this factor, thus measures only additive distortions.

3.4 Aggregation of disturbance in frequency and time

Following the concept that localised errors dominate perception [18, 24], PESQ integrates disturbance over several scales using a method designed to take optimal account of the distribution of error in time and amplitude. The disturbance values are aggregated using an *Lp* norm, which calculates a non-linear average using the following formula:

$$L_p = \left(\frac{1}{N} \sum_{m=1}^{N} disturbance[m]^p \right)^{1/p}$$

The disturbance is first summed across frequency using an *Lp* norm, giving a frame-by-frame measure of the perceived distortion. This frame disturbance is multiplied by two weightings. The first weight is inversely proportional to the instantaneous energy of the reference, raised to the power 0.04, giving slightly greater emphasis on sections for which the reference is quieter. This process replaces the silent interval weighting used in P.861. After this, the frame disturbance is bounded with an upper limit of 45. The second weight gives reduced emphasis on the start of the signal if the total length is over 16s, modelling the effect of short-term memory in subjective listening. This multiplies the frame disturbance at the start of the signal by a factor decreasing linearly from 1.0 (for files shorter than 16 seconds) to 0.5 (for files longer than 60 seconds).

After weighting, the frame disturbance is averaged in time over split second intervals of 20 frames (approx 320ms, accounting for the overlap of frames) using *Lp* norms. These intervals overlap 50%, and no window function is used. The split second disturbance values are finally averaged over the length of the

speech files, again using *Lp* norms. Thus the aggregation process uses three *Lp* norms – in general with different values of p – to map the disturbance to a single figure. The value of p is higher for averaging over the split second intervals to give greatest weight to localised distortions. The symmetric and asymmetric disturbance are averaged separately.

3.5 Realignment of bad intervals

In certain cases the first time alignment may fail to correctly identify a delay change, resulting in large errors for each section with incorrect delay. These are identified by labelling bad frames (which have a symmetric disturbance of more than 45) and joining together bad sections in which bad frames are separated by less than 5 good frames.

Each bad section is then realigned and the disturbance recalculated. Cross-correlation is used to find a new delay estimate. The auditory transform of the degraded signal is recalculated and the disturbance found. For each frame, if the realignment results in a lower disturbance value, the new value is used. Aggregation over split second intervals and the whole signal is performed after realignment.

3.6 MOS prediction and model calibration

To train PESQ a large number of different symmetric and asymmetric disturbance parameters were calculated by using different values of p for each of the three averaging stages. A linear combination of disturbance parameters was used as a predictor of subjective MOS. A further regression is required for each subjective test to account for context and voting preferences of different subjects, as discussed in section 4.1; for calibration a linear mapping was also used at this stage. Parameter selection was performed for all candidate sets of up to four disturbance parameters. The optimal combination – giving the highest average correlation coefficient – was found. This enabled the best parameters to be chosen from the full set of several hundred candidate disturbance parameters.

The partial compensation method used in PESQ means that it is not necessary to use many separate parameters to predict quality, for example to take account of filtering, modulation, or distribution of errors. Thus it was found that a combination of only two parameters – one symmetric disturbance and one asymmetric disturbance – gave a good balance between accuracy of prediction and ability to generalise. However, as this low-dimension model depends on earlier stages to incorporate complex perceptual effects, it was necessary to perform several design iterations. Coefficients in the auditory transform and disturbance processing were optimised then the optimal parameter combination was found, and the process repeated several times. The final training was performed on a database of 30 subjective tests.

The output mapping used in PESQ is given by

$$PESQMOS = 4.5 - 0.1\, disturbance_{SYMMETRIC} - 0.0309\, disturbance_{ASYMMETRIC}$$

For normal subjective test material the values lie between 1.0 (bad) and 4.5 (no distortion). In cases of extremely high distortion the *PESQMOS* may fall below 1.0, but this is very uncommon.

4. Performance results

4.1 Evaluation of performance

Condition MOS is the most common measure of subjective quality: this is the average MOS for four or more recordings of at least 8s in duration. These recordings are usually different sentence pairs spoken by two male and two female talkers; the condition MOS is therefore a material-independent measure of the quality of the connection. For comparison between objective and subjective score it is usual to compare the condition MOS with the condition average objective score.

However, a one-to-one comparison between objective and subjective MOS is not normally possible with tests conducted according to the ITU-T testing method [1, 2], because subjective votes are affected by factors such as the voting preferences of each subject or the balance of conditions in a test. This makes it impossible to directly compare results from one subjective test with another; some form of mapping between the two is required. The same is true for comparing objective scores with subjective MOS.

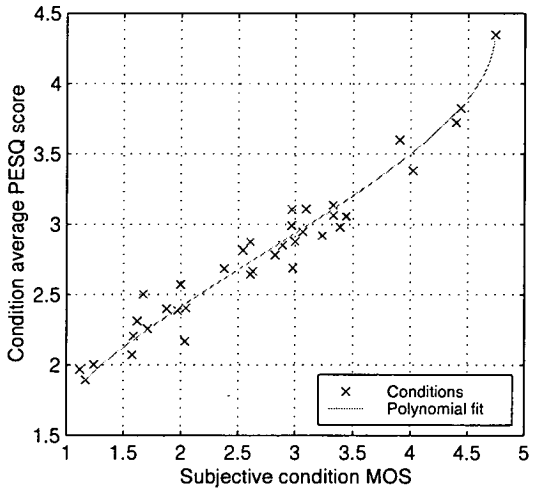
However, it is reasonable to expect that order should be preserved, so the difference between two sets of scores should be a smooth, monotonically increasing (one-to-one) mapping. The function used in ITU-T evaluation of objective models is a monotonic 3rd-order polynomial. This is applied, for each subjective test, to map the objective score onto the subjective score. It is then possible to calculate correlation coefficient and residual errors.

This process is illustrated by the following example, a subjective test on the performance of fixed and mobile networks with errors, noise and noise suppression. Figure 6(a) shows a scatter plot between subjective MOS and PESQ score, along with the monotonic 3rd-order polynomial fit with minimum mean squared error. The PESQ score is mapped by this polynomial to give a prediction of subjective quality, shown in Figure 6(b). The correlation coefficient for this test is 0.974, given by the following equation

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

where x_i is the condition MOS for condition i, $\bar{x}$ is the average of x_i across all conditions, y_i is the mapped condition-averaged PESQ score for condition i, and $\bar{y}$ is the average of y_i across all conditions.

(a) Raw scores and monotonic polynomial fit



(b) Scores after application of mapping

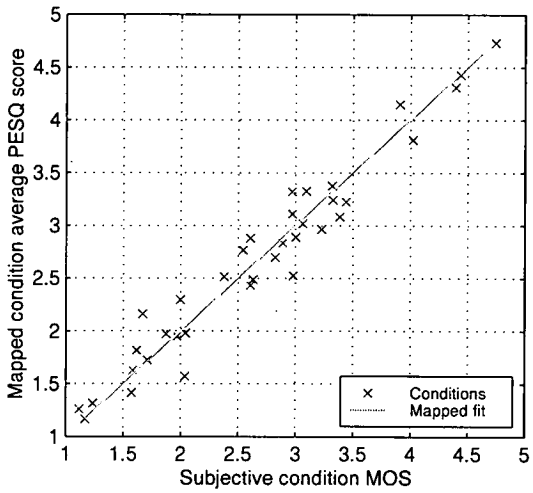


Figure 6: Mapping between objective and subjective MOS.

4.2 Correlation results

The performance of PESQ is compared to PSQM [12] and MNB [12 appendix II] in Figures 7–10 using correlations calculated according to the process described in the previous section. The figures plot the correlation coefficient between each model and subjective MOS for a number of ACR listening quality tests. Figure 7 presents 19 tests containing mainly mobile codecs and/or networks. Figure 8 gives results from 9 tests on predominantly fixed networks or codecs. Figure 9 shows 10 tests containing VoIP conditions on a wide range of codec/error types. Finally, Figure 10 gives the results for 8 tests conducted on PESQ by independent laboratories using data unknown in the development of the model. The tests

were conducted in a number of different languages, and 8 of the tests included conditions with background noise.

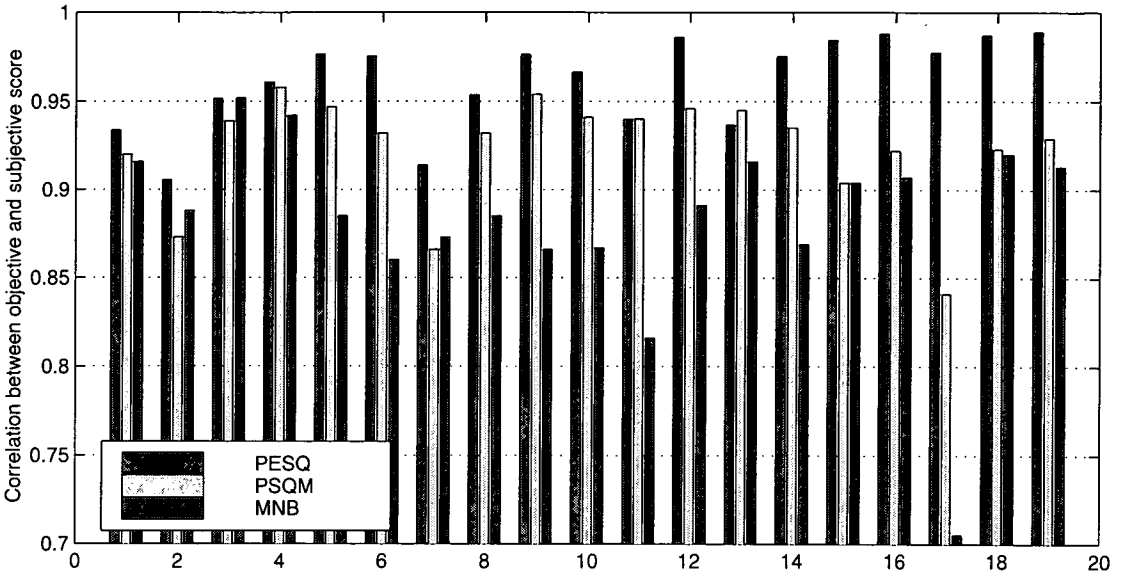


Figure 7: Mobile network performance results for PESQ, PSQM [12], and MNB [12]. Correlation coefficient, per condition, after monotonic 3rd-order polynomial mapping.

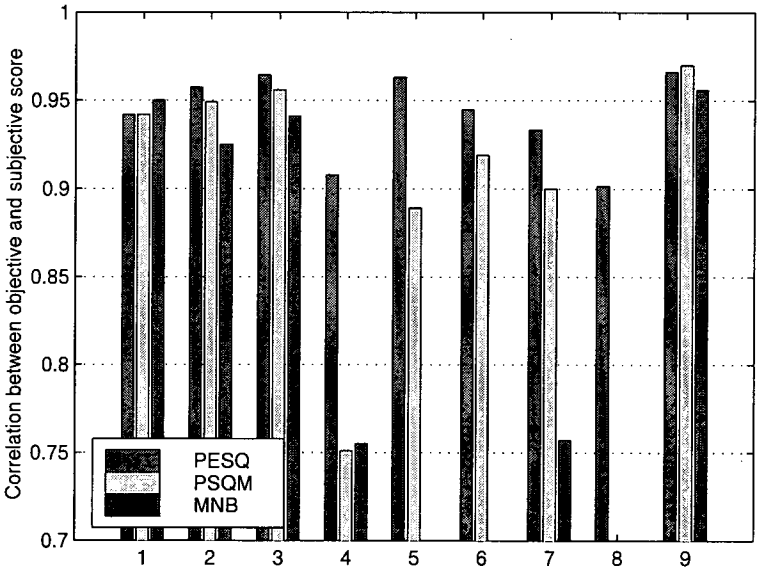


Figure 8: Fixed network performance results for PESQ, PSQM [12], and MNB [12]. Correlation coefficient, per condition, after monotonic 3rd-order polynomial mapping. In tests 5, 6 and 8 the scores for MNB (and PSQM in test 8) are off the bottom of the scale.

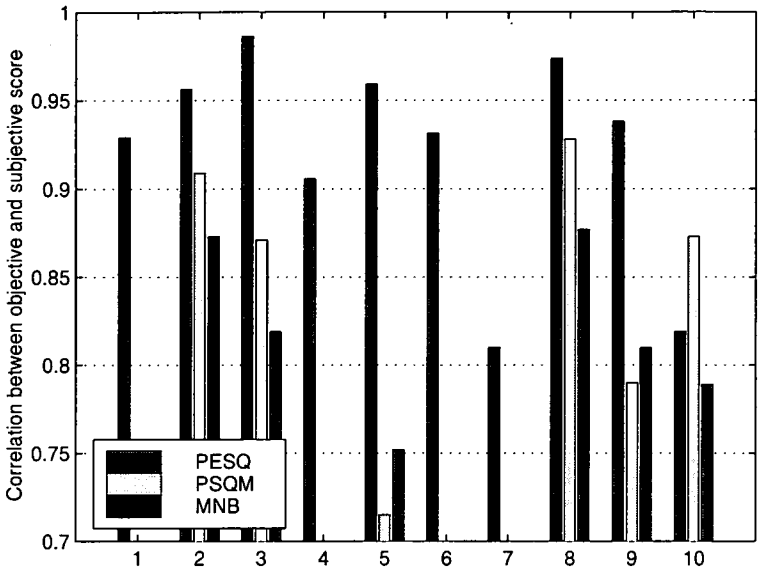


Figure 9: VoIP and multi-type test results for PESQ, PSQM [12], and MNB [12]. Correlation coefficient, per condition, after monotonic 3rd-order polynomial mapping. In tests 1, 4, 6 and 7 the scores for MNB and PSQM are off the bottom of the scale.

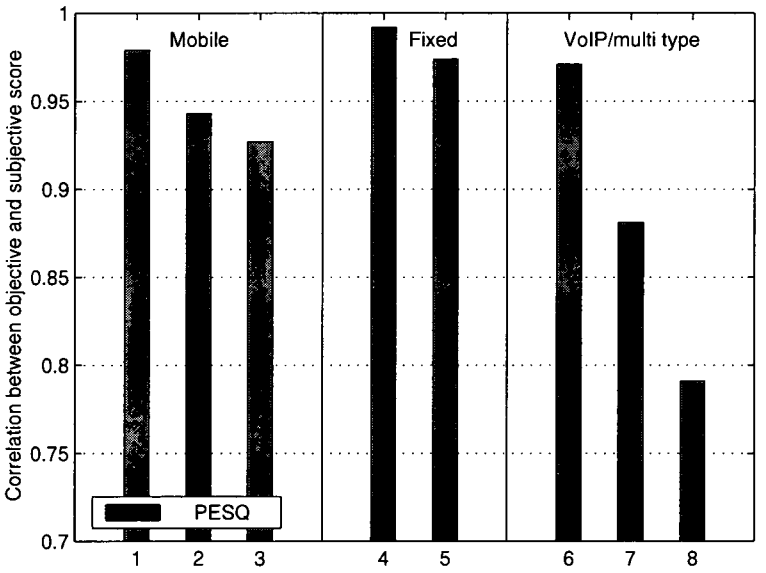


Figure 10: Independent results for unknown subjective tests (PESQ only). Correlation coefficient, per condition, after monotonic 3rd-order polynomial mapping.

4.3 Residual error distribution

A further method for measuring model performance is to plot the distribution of the absolute residual errors $|x_i - y_i|$ after the mapping described in section 4.1. Figures 11 plots the cumulative distribution of errors for PESQ, PSQM [12] and MNB [12 appendix II], calculated across 40 ACR listening quality tests containing a total of 1921 conditions. This shows, for example, that 93.5% of PESQ scores were within 0.5 MOS of the subjective score, and 100% of PESQ scores were within 1.125 MOS of the subjective score for these 40 tests.

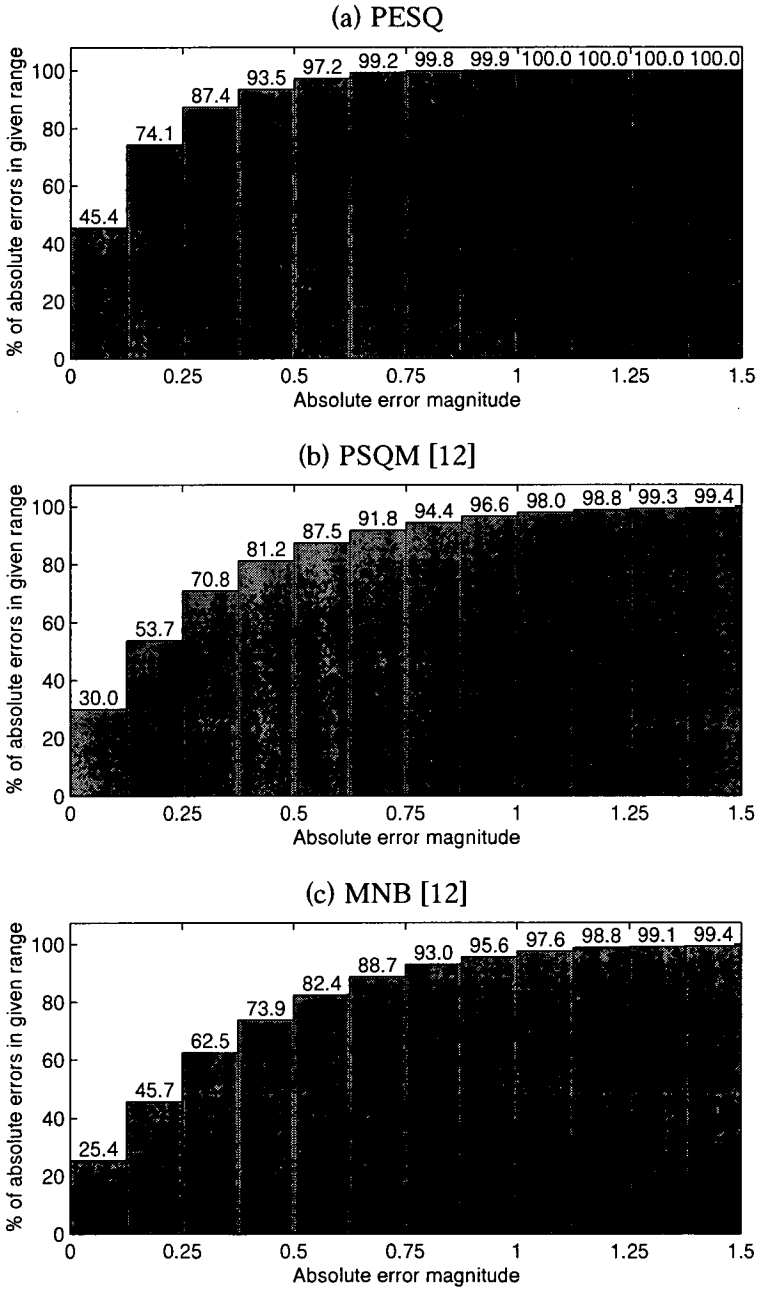


Figure 11: Residual error distribution for PESQ, PSQM [12], and MNB [12]. Per condition, after monotonic 3rd-order polynomial mapping.

5. Scope of PESQ

5.1 Conditions successfully tested

Table 4 presents a summary of the range of conditions for which PESQ has been tested and found to give acceptable performance. Full details of the scope of the model may be found in P.862 [21].

Test factors	Coding/network technologies	Measurement applications
Coding distortions Transmission/packet loss errors Multiple transcodings Environmental noise * Time warping (variable delay)	Waveform codecs (e.g. G.711, G.726, G.727) CELP/hybrid codecs at 4kbit/s and above (e.g. G.728, G.729, G.723.1) Mobile codecs and systems (e.g. GSM FR, EFR, HR, AMR; CDMA EVRC, TDMA ACELP, VSELP; TETRA)	Live network testing Network planning Codec evaluation/selection Equipment selection Codec/equipment optimisation

Table 4: Factors for which PESQ can be used for objective speech quality measurement.

* Note: for testing the effect of environmental noise, PESQ should be presented with the clean, unprocessed original and the noisy, coded, degraded signal.

5.2 Problems and areas for which PESQ is not applicable

PESQ is not intended to be used to assess:

- effect of listening level
- conversational delay
- talker echo/sidetone
- non-intrusive measurements.

Additionally, problems have been found with measurements on systems that replace speech with silence, for example front-end clipping or packet loss concealment with silence. See section 2.4 for more discussion of this.

5.3 Areas for further work

Certain applications of PESQ are currently under study or may require changes to the model, for example:

- wideband telephony/conferencing (16kHz sample rate)
- listener echo
- very low bit-rate speech vocoders (below 4kbit/s)
- head and torso simulator (HATS) measurements of handsets and/or hands-free telephones
- assessment of music.

One goal of further development is to extend the range of signal types and quality levels that a model can be used to assess. At present PESQ is calibrated to predict subjective tests conducted according to ITU-T P.800 or P.830 [1, 2] – i.e. "telephone quality", where subjects listen through a standard narrowband telephone handset. PEAQ [13, 14] is able to measure the quality of audio codecs – "audio quality" – for applications such as broadcast, with headphone or loudspeaker listening [3]. In between these two ranges is the so-called "intermediate quality" [23]. It is hoped that PESQ can be extended to provide assessment at both telephone and intermediate quality.

6. Summary and conclusion

PESQ performs much better than earlier codec assessment models such as P.861 PSQM and MNB, and is expected to replace them in early 2001 as a new ITU-T recommendation P.862. PESQ has been evaluated on a very wide range of speech codecs and telephone network tests. It has been found to produce accurate predictions of quality in the presence of diverse end-to-end network behaviours such as filtering and variable delay. PESQ represents a significant step forward in the accuracy and range of applicability of speech quality assessment models. It can be used for development, selection and optimisation of telephone network equipment and codecs, as well as for measurement applications such as network monitoring.

7. Acknowledgements

Thanks are due to ITU-T study group 12 question 13 for organising and driving the recent competition, and in particular the other proponents (Ascom, Deutsche Telekom and Ericsson) who contributed valuable test data and provided stiff competition. We would also like to thank the companies who acted as independent validation laboratories: AT&T, Lucent Technologies, Nortel Networks, and especially France Telecom R&D. We acknowledge the assistance of many of our colleagues at BT and KPN. Antony Rix is also supported by the Royal Commission for the Exhibition of 1851.

8. References

[1] Methods for subjective determination of transmission quality. ITU-T Recommendation P.800, August 1996.

[2] Subjective performance assessment of telephone-band and wideband digital codecs. ITU-T Recommendation P.830, August 1996.

[3] Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. ITU-R Recommendation BS.1116, July 1998.

[4] Schroeder, M. R., Atal, B. S. and Hall, J. L. "Optimizing digital speech coders by exploiting masking properties of the human ear". Journal of the Acoustical Society of America, 66 (6), 1647–1652, December 1979.

[5] Brandenburg, K. "Evaluation of quality for audio encoding at low bit rates". 82nd AES Convention, pre-print no. 2433, 1987.

[6] Karjalainen, J. "A new auditory model for the evaluation of sound quality of audio systems", IEEE ICASSP, 608–611, 1985.

[7] Beerends, J. G. and Stemerdink, J. A. "Measuring the quality of audio devices". 90th AES Convention, pre-print no. 3070, 1991.

[8] Beerends, J. G. and Stemerdink, J. A. "A perceptual audio quality measure based on a psychoacoustic sound representation". Journal of the AES, 40 (12), 963–974, December 1992.

[9] Beerends, J. G. and Stemerdink, J. A. "The optimal time-frequency smearing and amplitude compression in measuring the quality of audio devices". 94th AES Convention, pre-print no. 3604, 1993.

[10] Beerends, J. G. and Stemerdink, J. A. "A perceptual speech-quality measure based on a psychoacoustic sound representation". Journal of the AES, 42 (3), 115–123, March 1994.

[11] Beerends, J.G. "Measuring the quality of speech and music codecs, an integrated psychoacoustic approach". 98th AES Convention, pre-print no. 3945, 1995.

[12] Objective quality measurement of telephone-band (300–3400 Hz) speech codecs. ITU-T Recommendation P.861, February 1998.

[13] Method for objective measurements of perceived audio quality. ITU-R Recommendation BS.1387, January 1999.

[14] Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., Colomes, C., Keyhl, M., Stoll, G., Brandenburg, K. and Feiten, B. "PEAQ–The ITU standard for objective

measurement of perceived audio quality". Journal of the AES, 48 (1/2), 3–29, January/February 2000.

[15] Treurniet, W. C. and Soulodre, G. A. "Evaluation of the ITU-R objective audio quality measurement method". Journal of the AES, 48 (3), 164–173, March 2000.

[16] Wang, S., Sekey, A. and Gersho, A. "An objective measure for predicting subjective quality of speech coders". IEEE Journal on Selected Areas in Communications, 10 (5), 819–829, June 1992.

[17] Hollier, M. P., Hawksford, M. O. and Guard, D. R. "Characterisation of communications systems using a speech-like test stimulus", Journal of the AES, 41 (12), 1008–1021, December 1993.

[18] Hollier, M. P., Hawksford, M. O. and Guard, D. R. "Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain". IEE Proceedings – Vision, Image and Signal Processing, 141 (3), 203–208, June 1994.

[19] Rix, A. W., Reynolds, R. and Hollier, M. P. "Perceptual measurement of end-to-end speech quality over audio and packet-based networks". 106th AES Convention, pre-print no. 4873, May 1999.

[20] Rix, A. W. and Hollier, M. P. "The perceptual analysis measurement system for robust end-to-end speech quality assessment", IEEE ICASSP, June 2000.

[21] Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Draft Recommendation P.862, May 2000.

[22] Beerends, J.G. "Modelling cognitive effects that play a role in the perception of speech quality", in Proc. Int. Workshop on Speech Quality Assessment, Bochum, pages 1-9, November 1994.

[23] Rix, A. W. and Hollier, M. P. "Perceptual speech quality assessment from narrowband telephony to wideband audio", 107th AES Convention, pre-print no. 5018, September 1999.

[24] Quackenbush, S.R., Barnwell III, T.P., Clements, M.A. Objective measures of speech quality. Prentice Hall Advanced Reference Series, New Jersey USA, 1988.

Appendix D. Subjective test database

A total of 45 ACR listening quality subjective tests were used for this study. These are listed in the table, which shows the following.

- The experiment number in the author's database
- A brief description of the types of conditions in each test
- The laboratory that conducted the test
- The language of the speech material
- The number of reference, degraded file pairs in the test
- The number of distinct network conditions in the test
- Notes, using the key:
  - A Subjective test designed or conducted by the author
  - D Test containing variable-delay conditions
  - F Test containing more than one type of linear filtering
  - T Used in model training
  - V Used for model validation (Chapter 5)

Exp	Description	Source	Language	Files	Conds	Notes
1	GSM-FR, HR: bit errors and quantisation	BT	British	180	45	T
3	GSM-FR, HR: tandeming and bit errors	BT	British	180	45	V
4	GSM-FR, HR: bit errors and quantisation	BT	British	180	45	V
6	GSM-FR, HR: tandeming and bit errors	BT	British	180	45	V
8	Tetra: effect of tandeming, A-law	BT	British	180	45	T
9	Tetra: bit errors, A-law	BT	British	180	45	V
10	Tetra: bit errors, UPCM	BT	British	180	45	V
11	Tetra: bit errors, A-law	BT	British	180	45	V
12	Tetra: bit errors, UPCM	BT	British	180	45	V
13	G.729, G.728 and ADPCM codecs	BT	British	1152	48	V
14	GSM-FR, HR: bit errors, A-law	BT	British	2304	48	V
15	GSM-FR, HR: bit errors, UPCM, no IRS	BT	British	2304	48	V
16	GSM-FR, HR: tandeming, A-law/IRS and UPCM, no IRS	BT	British	2304	48	F V
17	G.729: interworking with other standard codecs	CNET	French	176	44	V
18	G.729: interworking with other standard codecs	NTT	Japanese	176	44	V
19	G.729: interworking with other standard codecs	BNR	American	176	44	T
23	G.729: effect of channel errors and noise	CNET	French	200	50	T
24	G.729: effect of channel errors and noise	CSELT	Italian	200	50	V
25	G.729: effect of channel errors and noise	NTT	Japanese	200	50	V

Exp	Description	Source	Language	Files	Conds	Notes
26	G.729: effect of channel errors and noise	BNR	American	200	50	V
27	AMR: static errors, clean, full rate, adaption off	BT	British	1128	47	T
28	AMR: static errors, clean, half rate, adaption off	BT	British	1128	47	V
29	AMR: static errors, clean, full rate, adaption on	BT	British	1128	47	V
30	AMR: static errors, clean, half rate, adaption on	BT	British	1128	47	V
31	BT P.86x proponent test: codecs, errors, transcodings, noise	BT	British	1200	50	A D F T
32	VoIP packet loss scenarios test	BT	British	3456	72	D V
34	Nortel objective model validation	Nortel	American	228	57	D
36	Fixed and mobile networks with noise and noise reduction	BT	British	144	36	A F T
38	DTX, frame/burst erasure, filtering and VAD test	BT	British	768	48	A F T
39	Berkom frame erasure test	DT	German	200	50	V
40	BT live VoIP measurement test	BT	German	200	50	A D T
41	Ascom proponent test 1	Ascom	French	116	29	V
42	Ascom proponent test 2: noise, VoIP and gain variation	Ascom	French	120	30	D V
43	Berkom proponent test: digital and acoustic tests with standard codecs	DT	German	200	50	F V
45	KPN proponent test: live VoIP networks	KPN	Dutch	60	60	D T
46	P86x Mobile codecs and background noise	BT	British	196	49	V
47	P86x Mobile codecs and background noise	DT	German	196	49	T
48	P86x ETSI VoIP measurement test	DT	German	208	52	D V
49	P86x Network Emulation: fixed network conditions	KPN	Dutch	200	50	A D T
50	P86x Network Emulation: fixed network conditions	Ascom	British	200	50	A D V
51	P86x Live Network Measurement	KPN	Dutch	184	46	D F V
53	P.SEAM Acoustic-Electric PSTN Handsets, headsets, HFT and network distortions	Psytechnics	British	648	54	F T
59	P.SEAM Acoustic-Electric mobile handsets, noise	DT/Opticom	German	600	50	F V
60	P.SEAM Acoustic-Electric mobile handsets, position	DT/ Swissqual	German	600	50	D F T
69	Recency effect, G.723.1 and packet loss	Psytechnics	British	400	20	A D T
45	**Totals**			25648	2119	

Appendix E. Example data

A CD-ROM accompanies this thesis and provides examples of some of the time-delay and frequency response estimation problems discussed in Chapter 3 and Chapter 4. To browse these examples, open the file index.html on the top level of the CD. The files are single-channel, and are provided at 16bits/sample, 8kHz sampling rate, in Windows .wav format (Intel byte order), with 44-byte headers.

References

Beerends 1992 — Beerends, J. G. and Stemerdink, J. A. A perceptual audio quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, 40 (12), 963–974, December 1992.

Beerends 1994 — Beerends, J. G. and Stemerdink, J. A. A perceptual speech-quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, 42 (3), 115–123, March 1994.

Beerends 1997 — Beerends, J. G. *Improvement of the P.861 perceptual speech quality measure,* ITU-T contribution COM12-20, December 1997.

Beerends 2000 — Beerends, J. G., Rix, A. W., Hekstra, A. P. and Hollier, M. P. *Proposed draft recommendation P.86x: Perceptual Evaluation of Speech Quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs.* ITU-T delayed contribution COM12-D140, April 2000.

Beerends 2002 — Beerends, J. G., Hekstra, A. P., Rix, A. W. and Hollier, M. P. Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part II – psychoacoustic model. *Journal of the Audio Engineering Society*, 50 (10), 765-778, October 2002.

Beerends 2003 — Beerends, J. G., Berger, J. and Rix, A. W. *Preliminary results for the P.AAM benchmark models.* ITU-T delayed contribution COM12-D109, January 2003.

Berger 1997 — Berger, J. *TOSQA – Telecommunication objective speech quality assessment,* ITU-T contribution COM12-34, December 1997.

Bernardo 2000 — Bernardo, J. M. and Smith, A. F. M. *Bayesian Theory.* John Wiley and Sons, March 2000.

Bishop 1995 — Bishop, C. M. *Neural Networks for Pattern Recognition.* Clarendon Press, November 1995.

Boucher 1981 — Boucher, R. E. and Hassab, J. C. Analysis of discrete implementation of generalized cross correlator. *IEEE Transactions on Acoustics, Speech and Signal Processing,* 29 (3), 609–610, June 1981.

Brandenburg 1987 — Brandenburg, K. *Evaluation of quality for audio encoding at low bit rates.* 82nd AES Convention, London, preprint 2433, March 1987.

Brandenburg 1996 — Brandenburg, K. and Bosi, M. Overview of MPEG audio: current and future standards for low-bit-rate audio coding. *Journal of the Audio Engineering Society*, 45 (1/2), 4–21, January 1996.

Brandstein 1995 — Brandstein, M. S. *A framework for speech source localization using sensor arrays.* Ph.D. thesis, Division of Engineering, Brown University, May 1995.

Brandstein 1997 — Brandstein M. S. and Silverman H. F. A practical methodology for speech source localization with microphone arrays. *Computer Speech and Language,* 11 (2), 91–126, 1997.

Bregman 1990 — Bregman, A. S. *Auditory scene analysis – the perceptual organisation of sound.* MIT Press, 1990.

Bronshtein 1985 — Bronshtein, I. N., Semendyayev, K. A. and Hirsch, K. A. *Handbook of mathematics.* Van Nostrand Reinhold, 1985.

Carlemalm 1997 — Carlemalm, C., Halvarsson, S. and Wahlberg, B. *Low complexity parameter estimation approach for fast time-delay estimation.* 36th IEEE Conference on Decision and Control, (2), 1603–1608, December 1997.

Carter 1981 — Carter, G. C. Guest editorial, time delay estimation (special issue). *IEEE Transactions on Acoustics, Speech and Signal Processing,* 29 (3), 461–462, June 1981.

Chan 1980 — Chan, Y. T., Riley, J. M. and Plant, J. B. A parameter estimation approach to time-delay estimation and signal detection. *IEEE Transactions on Acoustics, Speech and Signal Processing,* 28 (1), 8–16, February 1980.

Colomes 1995 — Colomes C., Lever M., Rault J.B. and Dehery Y.F. A perceptual model applied to audio bit-rate reduction. *Journal of the Audio Engineering Society*, 43 (4), 233-240, April, 1995.

Cooke 1993 — Cooke, M. *Modelling auditory processing and organisation.* Cambridge University Press, 1993.

Cooke 2001 — Cooke, M. and Ellis, D. P. W. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, 35, 141–177, 2001.

Cormen 1990 — Cormen, T. H., Leiserson, C. E. and Rivest, R. L. *Introduction to algorithms.* MIT Press, 1990.

Demuth 2001 — Demuth, H. and Beale, M. *Neural network toolbox for use with MATLAB.* Mathworks, 2001.

Duckworth 1968 — Duckworth, W. E. *Statistical techniques in technological research.* Methuen, 1968.

Feintuch 1981 — Feintuch, P. L., Bershad, N. J. and Reed, F. A. Time delay estimation using the LMS adaptive filter—Dynamic behavior. *IEEE Transactions on Acoustics, Speech and Signal Processing,* 29, 571–576, June 1981.

Friedman 1991 — Friedman, J. H. Multivariate adaptive regression splines. *Annals of Statistics,* 19, 1–141, March 1991.

Glasberg 1990 — Glasberg, B. R. and Moore, B. C. J. Derivation of auditory filter shapes from notched-noise data. *Hearing Research,* 47, 103–108, 1990.

GSM 06.10 — *European digital cellular telecommunications system (Phase 1); Full rate speech; Transcoding.* GSM 06.10, ETSI, February 1992.

GSM 06.20 — *European digital cellular telecommunications system; Half rate speech; Part 2: Half rate speech transcoding.* GSM 06.20, ETSI, December 1995.

GSM 06.60 — *Digital cellular telecommunications system (Phase 2+) (GSM); Enhanced Full Rate (EFR) speech transcoding.* GSM 06.60, ETSI, December 1997.

GSM 06.90 — *Digital cellular telecommunications system (Phase 2+) (GSM); Adaptive Multi-Rate (AMR) speech transcoding.* GSM 06.90, ETSI, December 1998.

GSM HR3 1993 — *Subjective Selection Tests (Phase 3) on the GSM Half-Rate Speech Coding Algorithm Candidate(s).* ETSI TCH-HS, TD 93/21 Issue 1.1, July 1993.

Handel 1989 — Handel, S. *Listening: an introduction to the perception of auditory events,* MIT Press, 1989.

Härmä 2000 — Härmä, A., Karjalainen M., Savioja, L., Välimäki, V., Laine, U. K. and Huopaniemi, J. Frequency-warped signal processing for audio applications. *Journal of the Audio Engineering Society,* 48 (11), 1011–1031, November 2000.

Hauenstein 1998 — Hauenstein, M. *Application of Meddis' inner hair-cell model to the prediction of subjective speech quality.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (1), 545–548, May 1998.

Haykin 1996 — Haykin, S. *Adaptive filter theory.* Prentice Hall, 1996.

Herre 2001 — Personal communication, February 2001.

Hollier 1993 — Hollier, M. P., Hawksford, M. O. and Guard, D. R. Characterisation of communications systems using a speech-like test stimulus. *Journal of the Audio Engineering Society*, 41 (12), 1008–1021, December 1993.

Hollier 1994 — Hollier, M. P., Hawksford, M. O. and Guard, D. R. Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain. *IEE Proceedings – Vision, Image and Signal Processing*, 141 (3), 203–208, 1994.

Hollier 1995 — Hollier, M. P. *Audio quality prediction for telecommunications speech systems.* Ph.D. thesis, University of Essex, 1995.

IEC 60651 — *Sound level meters.* IEC 60651, October 2001.

ISO 226 — *Acoustics – normal equal loudness contours.* ISO standard 226:1987.

ITU-R BS.1116 — *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems.* ITU-R Recommendation BS.1116, July 1998.

ITU-R BS.1387 — *Method for objective measurements of perceived audio quality.* ITU-R Recommendation BS.1387, January 1999.

ITU-T COM12-R2 — *Report of working party 2/12, Geneva, 9–16 April, 1997.* ITU-T Report COM12-R2, May 1997.

ITU-T G.107 — *The E-model, a computational model for use in transmission planning.* ITU-T recommendation G.107, July 2002.

ITU-T G.191 — *Software tools for speech and audio coding standardization.* ITU-T recommendation G.191, December 2000.

ITU-T G.711 — *Pulse code modulation of voice frequencies.* ITU-T recommendation G.711, 1988.

ITU-T G.723.1 — *Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s.* ITU-T recommendation G.723.1, March 1996.

ITU-T G.726 — *40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (ADPCM).* ITU-T recommendation G.726, 1990.

ITU-T G.728 — *Coding of speech at 16 kbit/s using low-delay code excited linear prediction.* ITU-T recommendation G.728, September 1992.

ITU-T G.729 — *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction.* ITU-T recommendation G.729, March 1996.

ITU-T P.48 — *Specification for an intermediate reference system.* ITU-T recommendation P.48, 1988.

ITU-T P.56 — *Objective measurement of active speech level.* ITU-T recommendation P.56, March 1993.

ITU-T P.58 — *Head and torso simulator for telephonometry.* ITU-T recommendation P.58, August 1996.

ITU-T P.800 — *Methods for subjective determination of transmission quality.* ITU-T recommendation P.800, August 1996.

ITU-T P.810 — *Modulated noise reference unit (MNRU).* ITU-T recommendation P.810, February 1996.

ITU-T P.830 — *Subjective performance assessment of telephone-band and wideband digital codecs.* ITU-T Recommendation P.830, August 1996.

ITU-T P.834 — *Methodology for the derivation of equipment impairment factors from Instrumental Models.* ITU-T Recommendation P.834, July 2002.

ITU-T P.861 — *Objective quality measurement of telephone-band (300–3400 Hz) speech codecs.* ITU-T Recommendation P.861, February 1998.

ITU-T P.862 — *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.* ITU-T Recommendation P.862, February 2001.

Johnson 1993 — Johnson, D. H. and Dudgeon, D. E. *Array signal processing: concepts and techniques.* Prentice Hall, 1993.

Karjalainen 1985 — Karjalainen M. *A new auditory model for the evaluation of sound quality of audio system.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Tampa, Florida, 608–611, March 1985.

Keogh 2001 — Keogh, E. J. and Pazzani, M. J. *Derivative dynamic time warping.* 1st SIAM International Conference on Data Mining, Chicago, April 2001.

Kitawaki 1988 — Kitawaki, N., Nagabuchi, H. and Itoh, K. Objective quality evaluation for low-bit-rate speech coding systems. *IEEE Journal on Selected Areas in Communications,* 6 (2), 242–248, February 1988.

Klaus 2000 — Klaus, H. *Report of the Question 13/12 Rapporteur's meeting (Solothurn, 6-10 March 2000).* ITU-T contribution COM12-117, March 2000.

Knapp 1976 — Knapp, C. H. and Carter, G. C. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing,* 24 (4), 320–327, August 1976.

Kubichek 1989 — Kubichek, R. F., Quincy, E. A. and Kiser, K. L. Speech quality assessment using expert pattern recognition techniques. *IEEE Pacific Rim conference on Communications, Computers and Signal Processing,* 208–211, June 1989.

Kubichek 1991 — Kubichek, R. F., Atkinson, D. and Webster, A. *Advances in objective voice quality assessment.* IEEE Global Telecommunications conference (Globecom), (3), 1765–70, December 1991.

Kubichek 1993 — Kubichek, R. F. Mel-cepstral distance measure for objective speech quality assessment. *IEEE Pacific Rim conference on Communications, Computers and Signal Processing,* (1), 125–128, May 1993.

Lagarias 1998 — Lagarias, J. C., Reeds, J. A., Wright, M. H. and Wright, P. E. Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on Optimization,* 9 (1), 112–147, 1998.

Lalou 1990 — Lalou, J. The information index: an objective measure of speech transmission performance. *France Telecom/CNET Annales des Télécommunications,* 45 (1-2), 47–65, 1990.

Liang 2001 — Liang, Y. J., Farber, N. and Girod, B. *Adaptive playout scheduling using time-scale modification in packet voice communications.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (3), 1445–1448, May 2001.

Lin 2001 — Lin, L., Ambikairajah, E. and Holmes, W. H. *Log-magnitude modelling of auditory tuning curves.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (5), 3293–3296, May 2001.

Ljung 1987 — Ljung, L. *System identification: theory for the user.* Prentice Hall, 1987.

Martin 1995 — Martin, K. D. *A computational model of spatial hearing.* M.Sc. thesis, Department of Electrical Engineering and Computer Science, MIT, 1995.

Mathworks 1998 — *MATLAB function reference, volume 1: language.* Mathworks Inc, 1998.

McHenry 1978 — McHenry, C. Multivariable subset selection. *Journal of the Royal Statistical Society, Series C,* 27 (23), 291-296, 1978.

Meky 1997 — Meky, M. M. and Saadawi, T. N. *Prediction of speech quality using radial basis functions neural networks.* 2nd IEEE Symposium on Computers and Communications, 174–178, July 1997.

Meyr 1984 — Meyr, H. and Spies, G. The structure and performance of estimators for real-time estimation of randomly varying time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing,* 32 (1), 81–94, February 1984.

Moore 1995 — Moore, B. C. J. (ed). *Hearing.* 2nd edition, 1995.

Moore 1997a — Moore, B. C. J. *An introduction to the psychology of hearing.* 4th edition, Academic Press, 1997.

Moore 1997b — Moore, B. C. J., Glasberg, B. R. and Baer, T. A model for the prediction of thresholds, loudness and partial loudness. *Journal of the Audio Engineering Society,* 45 (4), 224–239, April 1997.

Novorita 1999 — Novorita, B. *Incorporation of temporal masking effects into bark spectral distortion measure.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2), 665–668, March 1999.

Oppenheim 1989 — Oppenheim, A. V. and Schafer, R. W. *Discrete-time signal processing.* Prentice Hall, 1989.

Ordas 2001 — Ordas, P. and Fox, B. *Perceptual evaluation of speech quality – a discussion.* Microtronix Systems, http://microtronix.ca/pesq-disc.html, 2001.

Paillard 1992 — Paillard, B., Mabilleau, P., Morisette, S. and Soumagne, J. PERCEVAL: perceptual evaluation of the quality of audio signals. *Journal of the Audio Engineering Society*, 40 (1/2), 21–31, January 1992.

Park 2000 — Park, S. W., Ryu, S. K., Park, Y. C. and Youn, D. H. *A bark coherence function for perceived speech quality estimation.* International Conference on Spoken Language Processing (ICSLP), (2), 218-221, October 2000.

Patterson 1992 — Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C. and Allerhand, M. H. Complex sounds and auditory images. In Cazals, Y., Demany, L. and Horner, K. eds., *Auditory Physiology and Perception*, 429–446. Pergamon Press, 1992.

Quackenbush 1998 — Quackenbush, S. R., Barnwell III, T. P., Clements, M. A. *Objective measures of speech quality.* Prentice Hall, 1988.

Rajan 1997 — Rajan, J. J. and Rayner, P. J. W. Model order selection for the singular value decomposition and the discrete Karhunen–Loève transform using a Bayesian approach. *IEE Proceedings – Vision, Image and Signal Processing,* 144 (2), 116–123, April 1997.

Reed 1981 — Reed, F. A., Feintuch, P. L., and Bershad, N. J. Time delay estimation using the LMS adaptive filter—Static behavior. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29, 561–571, June 1981.

Reynolds 2001a — Reynolds, R. J. B. and Rix, A. W. Quality VoIP – an engineering challenge. *BT Technology Journal,* 19 (2), 23–32, April 2001.

Reynolds 2001b — Reynolds, R. J. B. and Rix, A. W. Achieving VoIP voice quality. In Swale, R. (ed), *Voice over IP: systems and solutions*, 29–49. IEE, December 2001.

Rix 1998a — Rix, A. W. and Hollier, M. P. *Robust design methodology for telephony assessment models.* ITU-T delayed contribution COM12-D031, February 1998.

Rix 1998b — Rix, A. W. and Hunt, T. J. *PAMS Trial Evaluation.* BT Systems Engineering document 421b02/T/007, March 1998.

Rix 1998c — Rix, A. W., Reynolds, R. J. B., Hollier, M. P., Sheppard, P. J. and Beamond, E. J. *Measurement of speech signal quality for networks exhibiting variable delay.* International patent application WO0022803, October 1998.

Rix 1998d — Rix, A. W., Beamond, E. J., Hollier, M. P. and Gray, P. *Performance metrics for objective quality assessment systems in telephony.* ITU-T delayed contribution COM12-D079, November 1998.

Rix 1998e — Rix, A. W. and Hollier, M. P. *Comparison of speech quality assessment algorithms: BT PAMS, PSQM, PSQM+ and MNB.* ITU-T delayed contribution COM12-D080, November 1998.

Rix 1999a — Rix, A. W., Bourret, A. and Hollier, M. P. Modelling human perception. *BT Technology Journal,* 17 (1), 24-34, January 1999.

Rix 1999b — Rix, A. W., Reynolds, R. J. B. and Hollier, M. P. *Perceptual measurement of end-to-end speech quality over audio and packet-based networks.* 106th AES Convention, Munich, preprint 4873, May 1999.

Rix 1999c — Rix, A. W. *Neural network training process.* European patent application EP1065601, July 1999.

Rix 1999d — Rix, A. W. *Comparison of opinion scales for subjective listening tests.* ITU-T delayed contribution COM12-D102, September 1999.

Rix 1999e — Rix, A. W. and Hollier, M. P. *Perceptual speech quality assessment from narrowband telephony to wideband audio.* 107th AES Convention, New York, preprint 5018, September 1999.

Rix 1999f — Rix, A. W., Reynolds, R. J and Hollier, M. P. *Robust end to end perceptual quality assessment of audio communication over packet-based networks.* IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'99), New Paltz NY, pp 39–42, October 1999.

Rix 2000a — Rix, A. W., Hekstra, A. P., Beerends, J. G. and Hollier, M. P. *Performance of the integrated KPN/BT objective speech quality assessment model.* ITU-T delayed contribution COM12-D136, April 2000.

Rix 2000b — Rix, A. W. and Hollier, M. P. *The perceptual analysis measurement system for robust end-to-end speech quality assessment.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Istanbul, (3), 1515-1518, June 2000.

Rix 2000c — Rix, A. W., Beerends, J. G., Hollier, M. P. and Hekstra, A. P. *PESQ – the new ITU standard for end-to-end speech quality assessment.* 109th AES Convention, Los Angeles, preprint 5260, September 2000.

Rix 2001a — Rix, A. W. *Results of quality assessment of wideband speech using PAMS.* ITU-T delayed contribution COM12-D001, January 2001.

Rix 2001b — Rix, A. W., Beerends, J. G., Hekstra, A. P. and Hollier, M. P. *Proposed modification to draft P.862 to allow PESQ to be used for quality assessment of wideband speech.* ITU-T delayed contribution COM12-D007, February 2001.

Rix 2002a — Rix, A. W. *A new PESQ-LQ scale to assist comparison between P.862 PESQ score and subjective MOS.* ITU-T delayed contribution COM12-D086, May 2002.

Rix 2002b — Rix, A. W., Hollier, M. P., Hekstra, A. P. and Beerends, J. G. Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part I – Time-delay compensation. *Journal of the Audio Engineering Society,* 50 (10), 755-764, October 2002.

Rix 2003a — Rix, A. W., Berger, J. and Beerends, J.G. *Perceptual quality assessment of telecommunications systems including terminals.* AES 114th Convention, Amsterdam, preprint 5724, March 2003.

Robert 1999 — Robert, A. *Results on perceptual invariants to transformations on speech.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (1), 393-396, 1999.

Roth 1971 — Roth, P. R. Effective measurements using digital signal analysis. *IEEE Spectrum*, 8, 62–70, April 1971.

Schroeder 1979 — Schroeder, M. R., Atal, B. S. and Hall, J. L. Optimizing digital speech coders by exploiting masking properties of the human ear. *Journal of the Acoustical Society of America*, 66 (6), 1647–1652, 1979.

Schroeder 1991 — Schroeder, J. E. and Kubichek, R. F. L1 and L2 normed cepstral distance controlled distortion performance. *IEEE Pacific Rim conference on Communications, Computers and Signal Processing,* (1), 41–44, May 1991.

Sekey 1984 — Sekey, A. and Hanson, B. Improved 1-bark bandwidth auditory filter. *Journal of the Acoustical Society of America,* 75 (6), 1902–1904, 1984.

Silverman 1990 — Silverman, H. F. and Morgan, D. P. The application of dynamic programming to connected speech recognition. *IEEE ASSP Magazine,* 7, 6–25, July 1990.

Söderström 1989 — Söderström, T. and Stoica, P. *System identification.* Prentice Hall, 1989.

Sporer 1997 — Sporer T. *Objective audio signal evaluation – applied psychoacoustics for modeling the perceived quality of digital audio.* 103rd AES-Convention, New York, preprint 4512, October 1997.

Stevens 1972 — Stevens, S. S. Perceived level of noise by Mark VII and decibels (E), *Journal of the Acoustical Society of America,* 51, 575–601, 1972.

Stuller 1997 — Stuller, J. A. and Hubing, N. New perspectives for maximum likelihood time-delay estimation. *IEEE Transactions on Signal Processing,* 45 (3), 513–525, March 1997.

Takahashi 1996 — Takahashi, A. and Kitawaki, N. *Review of validation tests for objective speech quality measures.* ITU-T contribution COM 12-74, May 1996.

Tallak 1993 — Tallak, S., Kubichek, R. and Schroeder, J. *Time delay estimation for objective quality evaluation of low bit-rate coded speech with noisy channel conditions.* IEEE Asilomar Conference, (2), 1216–1219, November 1993.

Tarraf 1999 — Tarraf, A. and Meyers, M. *Neural network-based voice quality measurement technique.* IEEE International Symposium on Computers and Communications, 375-381, 1999.

Thiede 1996 — Thiede, T. and Kabot, E. *A new perceptual quality measure for bit rate reduced audio.* 100th AES Convention, Copenhagen, preprint 4280, May 1996.

TIA/EIA IS-127 — *Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems.* TIA/EIA IS-127, July 1996.

Van Compernolle 1990 — Van Compernolle, D. Speech recognition in noisy environments with the aid of microphone arrays. *Speech Communication,* 9, 433-442, 1990.

Voiers 1977 — Voiers, W. D. *Diagnostic acceptability measure for speech communications systems.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 204–207, May 1977.

Voran 1997 — Voran, S. *Listener ratings of speech passbands,* IEEE Workshop on Speech Coding for Telecomms, 81–82, September 1997.

Voran 1999a — Voran, S. Objective estimation of perceived speech quality – part I: development of the measuring normalizing block technique. *IEEE Transactions on Speech and Audio Processing,* 7 (4), 371–382, July 1999.

Voran 1999b — Voran, S. Objective estimation of perceived speech quality – part II: evaluation of the measuring normalizing block technique. *IEEE Transactions on Speech and Audio Processing,* 7 (4), 383–390, July 1999.

de Vries 1994 — de Vries, D. K. *Identification of model uncertainty for control design.* Ph.D. thesis, Technische Universiteit Delft, September 1994.

Wang 1992 — Wang, S., Sekey, A. and Gersho, A. An objective measure for predicting subjective quality of speech coders. *IEEE Journal of Selected Areas in Communications*, 10 (5), 819–829, 1992.

Welch 1967 — Welch, P. D. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15 (2), 70–73, June 1967.

Yang 1997 — Yang, W., Dixon, M. and Yantorno, R. *A modified bark spectral distortion measure which uses noise masking threshold.* IEEE Workshop on speech coding for telecommunications, 55–56, September 1997.

Zwicker 1990 — Zwicker E. and Fastl H. *Psycho-acoustics, Facts and Models.* Springer Verlag, 1990.